Clock Synchronisation Assisted Clock and Data Recovery for Sub-Nanosecond Data Centre Optical Switching

A thesis submitted to UCL (University College London) for the partial fulfilment of the requirements for the degree of Doctor of Philosophy (PhD)

by

Kari Aaron Clark

Optical Networks Group
Department of Electronic and Electrical Engineering
UCL (University College London)

6th November 2020

Declaration

I, Kari Aaron Clark, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Copyright Notice

The way to deal with an impossible task was to chop it down into a number of merely very difficult tasks, and break each one of them into a group of horribly hard tasks, and each of them into tricky jobs, and each of them...

TERRY PRATCHETT, TRUCKERS

For Grandma Cats and Grandma Dogs.

Abstract

In current ‘Cloud’ data centres, switching of data between servers is performed using deep hierarchies of interconnected electronic packet switches. Demand for network bandwidth from emerging data centre workloads, combined with the slowing of silicon transistor scaling, is leading to a widening gap between data centre traffic demand and electronically-switched data centre network capacity. All-optical switches could offer a future-proof alternative, with potentially under a third of the power consumption and cost of electronically-switched networks. However, the effective bandwidth of optical switches depends on their overall switching time. This is dominated by the clock and data recovery (CDR) locking time, which takes hundreds of nanoseconds in commercial receivers. Current data centre traffic is dominated by small packets that transmit in tens of nanoseconds, leading to low effective bandwidth, as a high proportion of receiver time is spent performing CDR locking instead of receiving data, removing the benefits of optical switching. High-performance optical switching requires sub-nanosecond CDR locking time to overcome this limitation.

This thesis proposes, models, and demonstrates clock synchronisation assisted CDR, which can achieve this. This approach uses clock synchronisation to simplify the complexity of CDR versus previous asynchronous approaches. An analytical model of the technique is first derived that establishes its potential viability. Following this, two approaches to clock synchronisation assisted CDR are investigated: *1) Clock phase caching*, which uses clock phase storage and regular updates in a 2 km intra-building scale data centre network interconnected by single-mode optical fibre. *2) Single calibration clock synchronisation assisted CDR*, which leverages the 20× lower thermal sensitivity of hollow core optical fibre versus single-mode fibre to synchronise a 100 m cluster scale data centre network, with a single initial phase calibration step. Using a real-time FPGA-based optical switch testbed, sub-nanosecond CDR locking time was demonstrated for both approaches.

Keywords: *Clock Synchronisation Assisted Clock and Data Recovery*,
Clock Phase Caching, *Sub-Nanosecond Optical Switching*, *Hollow Core Fibre*,
Clock Synchronisation, *Data Centre Networks*, *All-Optical Data Centre Switching*.

Impact Statement

The lack of sub-nanosecond clock and data recovery in data centre optical switches was a key impediment to practical optical switching in the data centre, by limiting the achievable overall switching times, and therefore the achievable performance, of data centre optical switches. This thesis establishes clock synchronisation assisted clock and data recovery, an approach that leverages clock frequency and phase synchronisation to minimise clock and data recovery time to sub-nanosecond. This improvement in clock and data recovery time versus the previous state-of-the-art, by more than an order of magnitude, enables practical data centre all-optical switching.

In turn, all-optical switches could enable high-performance data centre networks that support emerging high-bandwidth hardware-based workloads such as deep neural network training, while simultaneously reducing data centre network power consumption by two thirds versus electronically-switched networks. Both outcomes are of extensive benefit to worldwide society: deep neural networks have many applications across fields including medicine, autonomous driving and machine translation; and minimising the power consumption of data centres, which could by 2030 consume 3 to 15% of global power, contributes towards limiting global warming.

The clock phase caching approach to clock synchronisation assisted clock and data recovery, which was investigated in collaboration with and patented by Microsoft Research, was published as a post-deadline conference paper at ECOC 2018, followed by a high-impact journal paper in Nature Electronics in 2020. Clock phase caching was used in Microsoft's prototype data centre all-optical switch, Sirius, presented at SIGCOMM 2020, in which the sub-nanosecond clock and data recovery locking time of clock phase caching was a key enabler of its high performance.

The single calibration approach to clock synchronisation assisted clock and data recovery using the low thermal sensitivity of hollow core fibre, investigated in collaboration with the Optoelectronics Research Centre, was published as a top-scored conference paper at ECOC 2019, followed by an invited, highly-scored journal paper in JLT in 2020. This work was also, to my and my coauthor's knowledge, the first demonstration of optical switching using hollow core fibre.

The analytical modelling of clock synchronisation assisted clock and data recovery established in this thesis, which explains why the approach works, is not yet published, and is intended to lead to a further one to two journal papers.

The work also led to two UK national competitions wins: *1)* The Connected Nation Pioneers 2018 Competition, in which I was Overall Winner and Winner of the Intelligent Informatics category, against over 100 other PhD student competitors. *2)* STEM for BRITAIN 2019, held at UK Parliament, in which I won Bronze in Engineering, against over 30 other junior researcher competitors in my category.

Clock synchronisation assisted clock and data recovery is also likely to have further applications beyond data centre optical switching, for instance in optical access networks. The distributed phase synchronisation enabled by the approach could also be of value in many applications that require synchronisation, such as quantum key distribution clock networks, and in the optical backbone supporting fifth-generation (5G) wireless networks.

Acknowledgements

For many reasons, the last six years of my life have been exceedingly difficult. While that is true, the last six years of my life have also been highly rewarding, and highly fulfilling. Many difficult challenges were overcome in the process of studying for my PhD, both academic and personal, and there are many to whom I owe thanks.

For the last three years of my PhD, I have been supervised by Dr. Zhixin Liu. His extensive knowledge of optics and analog electronics, wealth of excellent general advice, and infectious enthusiasm for research, has been invaluable. I am delighted that many of the ideas and goals that we discussed together – some of them rather ambitious – did with time and a lot of effort come to fruition. I also thank Dr. Philip Watts, who supervised me for the first two years of my PhD. His expertise on optical switching got my project started, and I owe my FPGA programming expertise to him. Prof. Polina Bayvel, Prof. Izzat Darwazeh and Dr. Georgios Zervas also provided invaluable academic support and advice at many points throughout my project.

Microsoft Research Cambridge have given much to the project. In addition to supporting my PhD financially through the Microsoft Optics for the Cloud programme, their industrial perspective on data centre networks acted to strengthen the depth and impact of my project. Dr. Benn Thomsen, Dr. István Haller, Dr. Krzysztof Jozwik and Mr Hugh Williams all stand out, but I am particularly thankful for the academic advice, supervision and enthusiasm of Dr. Paolo Costa, closely followed by Dr. Hitesh Ballani, who have both particularly closely supported my PhD throughout. I am also grateful for the opportunity I was given to complete a 9-month internship at Microsoft Research Cambridge in 2016, which I thoroughly enjoyed.

The Optoelectronics Research Centre (ORC) at the University of Southampton allowed me to add a new dimension to my PhD by granting me the opportunity to investigate the impact of using their low thermal sensitivity hollow core fibre on clock synchronisation assisted clock and data recovery. The ORC also supported my PhD financially through the EPSRC Lightpipe Project. Dr. Yong Chen, Dr. Eric R. Numkam Fokua, Dr. Tom Bradley, Prof. Francesco Poletti and Prof. David J. Richardson all contributed, but I extend particular thanks to Prof. Radan Slavík for his academic advice, support and fantastic humour.

I also extend my appreciation to Inphi Corporation, who funded the first three years of my PhD through the UCL Impact Award. Mr Mike Harwood and Mr Andre Szczepanek provided excellent industrial advice and support throughout these first three years, and their extensive experience with designing electronic high-speed serial transceivers, along with their industrial view on data centre interconnects, was particularly valuable – and I also rather enjoyed the academic discussions assisted by the pints of cider they bought me at their excellent local pub in Northampton.

I have greatly enjoyed studying alongside my fellow PhD students and friends in the Optical Networks Group. In particular, (now) Dr. Paris Andreades, (now) Dr. Joshua Benjamin, (now) Dr. Thomas Gerard and Mr Ronit Sohanpal stand out. Our many discussions about optical interconnects have been highly stimulating, and it has been a pleasure to work together with them towards enabling optically-switched data centres.

Outside of work, the many board games evenings and other events that I attended (and helped organise, first as Society Treasurer for a year, and later as Webmaster for two years) at the UCL Science-Fiction and Fantasy Society have contributed much to keeping me sane during stressful times. There are many names – too many to include here – but I am particularly thankful for the support of Miss Andrea Petrou, who was instrumental in helping me get through one of the most difficult periods of my PhD.

I am, as always, grateful for the love and support of my family: Mum, Dad, Laif, Vena, Grandma Cats and Grandma Dogs. They, as they always have done, helped me to succeed in my pursuits, both academic and personal, during my PhD. There have been many very difficult points during the last six years, and without their support, it is likely that I would not have completed my PhD. I do wish that my maternal grandmother, nicknamed Grandma Cats within my family, who died during the second year of my PhD, had lived to see me fulfil my dream of contributing something new for science and engineering – to place my brick in the wall – but nonetheless, I am sure that she, like the rest of my family, would have been proud of me.

Finally, I would like to thank my partner, Elisabeth (E.V.V.), who has been consistently by my side, and supported me, during the latter part of my PhD. From exploring London, to doing photography in the Scottish Highlands, to writing our PhD theses together under the COVID-19 lockdown, she has made my life, and by extension also my work, brighter, and better.

Table of Contents

List of Figures

List of Tables

List of Terms and Abbreviations

ASHRAE American Society of Heating, Refrigerating and Air-Conditioning Engineers 63

ASIC application-specific integrated circuit 35, 39, 49, 52, 80, 82, 168

AWG arrayed waveguide grating 35, 165, 166, 171

AWGR arrayed waveguide grating router 58, 59

BB-PD bang-bang phase detector 16, 69–71, 128, 149

BER bit error rate 20, 38, 53, 56, 67, 68, 74, 90, 118, 127, 170, 173, 174, 177, 193–198, 238

CDR clock and data recovery 16, 18–20, 22, 37–45, 56–58, 60, 61, 65–73, 75, 79–81, 96, 116, 122, 125–134, 136–139, 141, 143, 144, 146–151, 156–161, 163, 165–167, 169–181, 183, 185–187, 189, 191, 192, 194–199, 201, 203–207, 237–240

CID consecutive identical digits 67

CPU central processing unit 52

CRAC computer room air conditioner 62

CRC cyclic redundancy check 237, 238, 240

CSA-CDR clock synchronisation assisted clock and data recovery 17, 18, 20, 21, 42–44, 79–83, 116, 122, 124–127, 138, 146, 150–154, 160, 165, 167, 168, 185, 186, 189–191, 193–198, 202, 204–206, 208, 235–237, 240, 242, 243

CTLE continuous time linear equaliser 166

CW continuous wave 165, 193

CWDM coarse wavelength division multiplexing 201, 207

DAC digital to analog converter 67

DDR3 SDRAM Double Data Rate 3 Synchronous Dynamic Random-Access Memory 240

DFB distributed feedback 193

DMT discrete multi-tone 56

DSP digital signal processing 53

DTP Datacentre Time Protocol 17, 77

DWDM Dense Wavelength Division Multiplexing 56

EAM electro-absorption modulator 164, 193

EDFA Erbium-doped fibre amplifier 165, 166, 194

EML externally modulated laser 164, 193

FEC forward error correction 53, 56, 100, 170

FIFO first-in first-out queue 235, 238, 239

FPGA field programmable gate array 21, 42, 43, 56, 69, 78, 109, 121, 124, 147, 151, 163–169, 179, 189, 191, 193, 195–197, 235–240

FTTH fibre to the home 21, 205, 206

GPS Global Positioning System 75

GPU graphical processing unit 52

GVCO gated voltage controlled oscillator 16, 66, 67, 72, 203

HC-NANF hollow-core nested anti-resonant nodeless fiber 201

HC-PBGF hollow-core photonic bandgap fibre 20, 21, 188, 194, 195, 197, 198, 200–202, 241–243

HCF hollow core fibre 20, 21, 43, 45, 80, 124, 150, 184–202, 204, 207

ILA integrated logic analyser 240

IM-DD intensity modulation direct detection 56

ISI intersymbol interference 83, 94, 96, 123

JTAG Joint Test Action Group 240

LC-Tank inductor-capacitor tank 179, 180

LHC Large Hadron Collider 78

MAC media access control 169

MCF multi-core fibre 124, 150, 207

MEMS microelectro-mechanical system 57, 58, 61

MFD mode field diameter 201

MMF multi-mode fibre 49, 54–56

MZI Mach-Zehnder interferometer 37, 58, 59

MZM Mach-Zehnder modulator 164, 165, 172, 179, 181, 193, 194

NRZ non-return to zero 54–56, 90–95, 97–100, 107, 109, 112, 113, 121–123, 133, 135, 143, 189, 193, 228, 229

NRZ-OOK non return to zero on off keying 17, 18, 22, 42, 49, 54–56, 83, 84, 87, 97, 99–101, 103–112, 114, 115, 117–123, 182, 190, 194, 207

OEO optical-electrical-optical 59

OOK on-off keying 17, 18, 99, 101, 102, 110, 114, 117, 119, 121, 135, 136, 143, 164

OSNR optical signal-to-noise ratio 166

OTDM optical time division multiplexing 16, 44, 73–75

PAM-4 4-level pulse amplitude modulation 49, 54–56, 121, 207

PAM-8 8-level pulse amplitude modulation 56, 207

PDF probability distribution function 111, 230

PI phase interpolator 16, 69–71, 79, 149, 166, 203, 237, 238

PIC photonic integrated circuit 60

PIN p-i-n junction 17, 22, 83, 97, 99–108, 110, 120–123, 135, 143, 182, 190

PIPPM phase interpolator parts per million 167, 168

PISO parallel in serial out 167, 237

PLL phase locked loop 66–68, 70, 167, 179, 180, 193, 237

PON passive optical network 45, 68

ppm parts per million 69, 73

PRBS pseudo random binary sequence 67, 169, 193, 237–239

PTP Precision Time Protocol 17, 77, 78

PVT process voltage temperature 65, 67

QKD quantum key distribution 208

QPLL quad phase locked loop 179, 237

RMS root mean square 179–182

RZ return to zero 74, 75

SDH Synchronous Digital Hierarchy 17, 75, 76

SDM space division multiplexing 56, 207

SERDES serialiser-deserialiser 77, 78

SFP Small Form Factor Pluggable 78

SIPO serial-in parallel-out 238

SMF-28 single-mode fibre 17, 20, 43–45, 49, 54–56, 62, 80, 81, 85–87, 123, 124, 147, 159, 161, 164, 166, 171, 172, 175, 178, 182, 183, 185–189, 191, 193, 196, 197, 199–201, 204, 241–243

SOA semi-conductor optical amplifier 35, 37, 58–60

SONET Synchronous Optical Networking 17, 75, 76

SWDM shortwave wavelength division multiplexing 56

Sync-E Synchronous Ethernet 17, 75, 76, 78

TCD thermal coefficient of delay 86, 88, 186–189, 191, 201

TDC temperature delay coefficient 186

TDM-PON time division multiplexed passive optical network . . 21, 205, 206, 208

TIA transimpedance amplifier . 83, 97, 100–107, 109, 120–123, 135, 143, 144, 165, 190, 193

ToR top-of-rack 45, 49, 50, 52, 54, 63, 79, 124, 150

UART universal asynchronous receiver transmitter 238–240

UTC Coordinated Universal Time 75

VCO voltage controlled oscillator 66, 67, 72

VCSEL vertical-cavity surface-emitting laser 56

VIO virtual input/output 240

WDM wavelength division multiplexing 74

List of Symbols

L	number of Clos-switch hierarchy layers
k	number of bidirectional ports per electronic switch
t	time, time-of-flight
n_g	group refractive index
c	speed of light in vacuum
T	absolute temperature
L	fibre length, distance between nodes
$\frac{dT}{dt}$	rate of change of temperature
Δt	change in fibre time of flight
ΔT	change in temperature
τ	fibre thermal coefficient of delay
ϕ, α, β	clock phase
B	symbol rate
$\triangleq$	defined as equal to
$\Delta\phi$	change in clock phase
$\omega(\phi)$	ideal square NRZ pulse
f	frequency
f_c	cut-off frequency (-3 dB frequency)
$\hat{H}(f)$	amplitude normalised Gaussian filter frequency response
$\hat{f}(t)$	amplitude normalised Gaussian filter impulse response
σ_t	Gaussian filter impulse response standard deviation in terms of time
FWHM_t	Gaussian filter impulse response full width half maximum in terms of time

Δt_{20-80}	rise time, 20% to 80%
Δt_{80-20}	fall time, 80% to 20%
k_{pls}	Gaussian filter impulse response in terms of symbols
FWHM_k	Gaussian filter impulse response full width half maximum in terms of symbols
$v(\phi)$	Gaussian filtered NRZ pulse shape in terms of symbols
$\text{erf}(z)$	error function of z
$\text{erfc}(z)$	complementary error function of z
$v_0(\phi)$	positive-going Gaussian NRZ pulse in terms of symbols
$v_1(\phi)$	negative-going Gaussian NRZ pulse in terms of symbols
v_{peak}	peak Gaussian NRZ pulse amplitude
$(v_1 - v_0)_{\text{peak}}$	peak Gaussian NRZ eye height
p_e	bit error probability (expectation of the bit error rate)
$p(0)$	probability of receiving a 0
$p(1)$	probability of receiving a 1
$P(0\|1)$	probability of erroneously sampling a 0 instead of a 1
$P(1\|0)$	probability of erroneously sampling a 1 instead of a 0
$p_e(\phi)$	bit error probability as a function of clock phase offset
$P(0\|1,\phi)$	probability of erroneously sampling a 0 instead of a 1 as a function of clock phase offset
$P(1\|0,\phi)$	probability of erroneously sampling a 1 instead of a 0 as a function of clock phase offset
$I(\phi)$	sampled photocurrent as a function of sampling clock phase offset
$I_D(\phi)$	photocurrent decision threshold as a function of sampling clock phase offset
$I_0(\phi)$	mean photocurrent of the positive-going Gaussian NRZ pulse as a function of sampling clock phase offset
$I_1(\phi)$	mean photocurrent of the negative-going Gaussian NRZ pulse as a function of sampling clock phase offset

$\sigma_0(\phi)$	standard deviation of the photocurrent of the positive-going Gaussian NRZ pulse as a function of sampling clock phase offset
$\sigma_1(\phi)$	standard deviation of the photocurrent of the negative-going Gaussian NRZ pulse as a function of sampling clock phase offset
I	photocurrent
R	PIN photodiode responsivity
P_{opt}	incident optical power
I_{max}	maximum photocurrent
I_{avg}	average photocurrent
I_{min}	minimum photocurrent
P_{max}	maximum incident optical power
P_{avg}	average incident optical power
P_{min}	minimum incident optical power
r_e	extinction ratio
σ_T	standard deviation of the photodiode thermal noise
σ_d	standard deviation of the photodiode dark current noise
σ_s	standard deviation of the photodiode shot noise
σ_{ni}	standard deviation of the TIA input referred noise
k_B	Boltzmann's constant
q	charge of a single electron
F_n	TIA noise figure
R_L	TIA transimpedance
I_p	photodiode photocurrent
I_d	photodiode dark current
σ^2	total PIN photoreceiver noise variance
σ_T^2	PIN photoreceiver thermal noise variance
σ_d^2	PIN photoreceiver dark current noise variance
σ_{opt}^2	PIN photoreceiver optical power noise variance

$\sigma_0^2(\phi)$	variance of the photocurrent of the positive-going Gaussian NRZ pulse as a function of sampling clock phase offset
$\sigma_1^2(\phi)$	variance of the photocurrent of the negative-going Gaussian NRZ pulse as a function of sampling clock phase offset
$p_{e(\text{post-jit})}$	bit error probability, including the impact of jitter
$p_{e(\text{pre-jit})}$	bit error probability, before including the impact of jitter
PDF_{jit}	probability density function of the jitter
σ_{jit}	standard deviation of the jitter in terms of time
k_{jit}	standard deviation of the jitter in terms of symbols
$p_e(t)$	bit error probability as a function of time since the beginning of data packet reception
$p_{e(\text{overall})}(t)$	overall packet bit error probability
t_{lock}	CDR locking time
$p_{e(\text{lock})}$	bit error probability threshold at which CDR locking is defined as complete
$t_{\text{pkt-len}}$	length of one data packet
$\Phi_{\text{ideal}}(\phi)$	ideal clock phase error of a bang-bang phase detector
$\Phi_{\text{jittered}}(\phi)$	jittered clock phase error of a bang-bang phase detector
$\in$	is a member of
$\mathbb{Z}$	the set of all integers
$\text{E}[z]$	expectation of z
ϕ_0	initial clock phase offset
$d\phi$	small clock phase correction applied by a clock phase interpolator CDR
dt	clock phase interpolator CDR clock phase offset measurement interval
p	clock phase interpolator CDR proportional gain constant
$\phi(t)$	clock phase offset as a function of time since the beginning of data packet reception

$I_0(t)$	mean photocurrent of the positive-going Gaussian NRZ pulse as a function of time since the beginning of packet reception
$I_1(t)$	mean photocurrent of the negative-going Gaussian NRZ pulse as a function of time since the beginning of packet reception
$k_{\text{jit-h}}$	high frequency jitter standard deviation
$k_{\text{jit-l}}$	low frequency jitter standard deviation
$\text{PDF}_{\text{jit-h}}$	probability density function of the high frequency jitter
$\text{PDF}_{\text{jit-l}}$	probability density function of the low frequency jitter
$p_{e(\text{post-jit-h})}(t)$	bit error probability as a function of time since the beginning of packet reception, including the impact of high frequency jitter
$p_{e(\text{post-jit-l})}(t)$	bit error probability as a function of time since the beginning of packet reception, including the impact of low frequency jitter
$p_{e(\text{post-jit})}(t)$	bit error probability as a function of time since the beginning of packet reception, including the impact of all jitter frequencies
t^*	time elapsed since last calibration of the clock phase offset (over time scales of magnitude much much greater than t_{pkt})
$\Delta T(t^*)$	change in temperature as a function of time since the last calibration of the clock phase offset
f_ϕ	clock phase update rate
$\phi_{\min}$	minimum initial clock phase offset at the beginning of packet reception
$\phi_{\max}$	maximum initial clock phase offset at the beginning of packet reception
$p_{e(\text{ph-cached})}(t)$	clock phase cached bit error probability as a function of time since the beginning of packet reception
$\phi_{\max\,(\text{thresh})}$	threshold maximum initial clock phase offset at which bit error probability rises above a defined acceptable threshold
$f_{\phi(\text{min})}$	minimum required clock phase update rate
k	proportionality constant between rate of change of temperature and the minimum required clock phase update rate
o	throughput overhead from clock phase caching
N	total number of nodes (servers / switches) connected to an optical switch

n	Tx-Rx node pair iterator
$f_{\phi(n)}$	clock phase update rate for Tx-Rx node pair n
t_{meas}	time required at a single receiver to perform a single clock phase update measurement
t_{update}	time required at a single transmitter to send a single clock phase update value
τ_n	fibre thermal coefficient of delay for Tx-Rx node pair n
L_n	distance between nodes for Tx-Rx node pair n
B_n	symbol rate for Tx-Rx node pair n
$\tau_{\max(\text{thresh},n)}$	threshold maximum initial clock phase offset at which bit error probability rises above a defined acceptable threshold for Tx-Rx node pair n
$\left(\frac{dT}{dt}\right)_n$	rate of change of temperature for Tx-Rx node pair n
$o_{\max}$	worst-case throughput overhead from clock phase caching
$\Delta T_{\max}$	upper bound on the $<$625 ps (under 16 symbol) CDR locking time temperature window
$\mathcal{F}^{-1}\{z\}$	inverse Fourier transform of z

Chapter 1

Introduction

1.1 Scope of the Thesis

GLOBAL INTERNET TRAFFIC is increasing rapidly, driven primarily by demand from consumers for high bandwidth content such as video traffic [3]. To minimise the use of expensive and high latency inter-continental links, content is increasingly replicated throughout the world and delivered locally from regional data centres. To maximise the energy and cost efficiency of these regional data centres, each data centre typically contains hundreds of thousands of servers, which are interconnected by deep 3 to 5-layer hierarchies of electronic switch application-specific integrated circuits (ASICs). Growth in consumer and business demand, combined with emerging hardware workloads such as the training of deep neural networks and distributed key value stores, is driving a 100% increase of network traffic within these data centres every 12 to 18 months [4].

Concurrently, the growth rate of electronic switch ASIC bandwidth is slowing [5], due to the nearing of fundamental physical limits on the achievable power consumption and density of silicon transistors [6]. The impact of this at network scale is that to continue to meet current growth in data centre traffic, increasingly deep hierarchies of electronic switches are required, as shown in Figure 1.1a, with increasingly prohibitive cost, power consumption and latency [7, 8, 9]. For example, a current large data centre would ideally be interconnected by a 100 Pb/s non-blocking network, which would consume 48.7 MW of power, which is more than the 32 MW that is typically allocated for the entire data centre [8, 9]. Consequently, as the scaling of electronic switch ASICs continues to slow, and the gap between the capacity of electronically switched networks and data centre traffic increases, alternative solutions to data centre networking will be required that are less constrained by silicon transistor scalability.

All-optically-switched data centre networking, as shown in Figure 1.1b, is a potential candidate for overcoming this challenge. In all-optical switching, momentary light paths are created to send data packets (64-1500 Bytes [10]) directly between end-points (such as servers or electronic switches) using optical switching elements such as semi-conductor optical amplifiers (SOAs), arrayed waveguide gratings

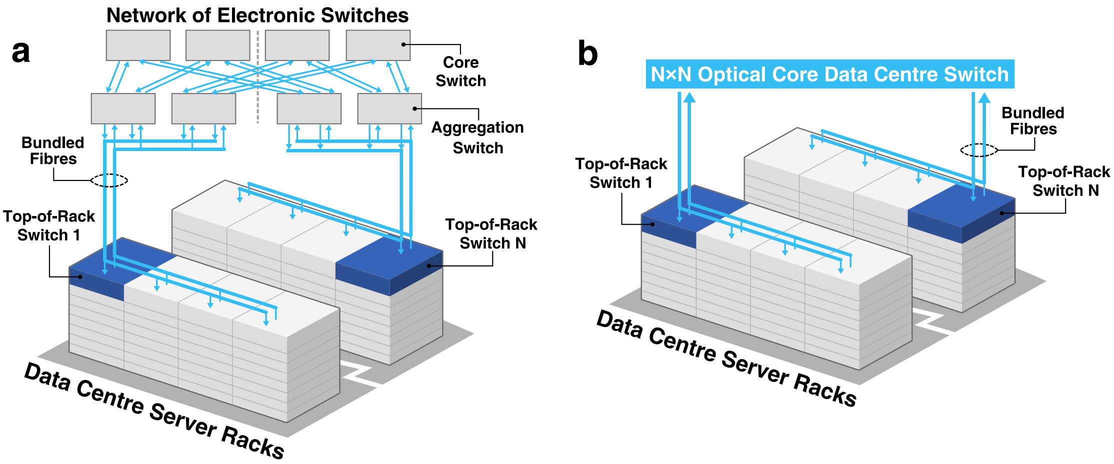

Fig. 1.1: Data centre electronically-switched and all-optically-switched network architectures. **a**, Current electronically-switched data centre network architecture, which interconnects servers using a deep hierarchy (4 to 5 levels, 3 shown here) of hundreds to thousands of small bandwidth electronic switches. **b**, A potential future all-optically-switched data centre architecture, which would use a single large port count $N \times N$ optical switch to interconnect all top-of-rack switches. All-optically-switched architectures promise large improvements in latency, power consumption and capital cost versus the current electronically-switched approach [8, 9].

(AWGs) and tuneable lasers. These optical paths are only established for very short lengths of time due to the short transmission time of data centre packets, e.g. 20 ns for a 64 Byte minimum size data packet at 25 Gb/s [10]. Data centre network simulations of large-port count all-optical switches have shown that they can closely match the performance of an ideal non-blocking 100 Pb/s electronically-switched network, while consuming 74 to 77% less power [8] and costing 82% less to build [8]. Even when comparing a non-blocking all-optically switched network with an equivalent current oversubscribed electronically-switched network, all-optically switched networks can have 40% less power consumption [9] and cost 47% less to build [8], while having 100× smaller worst-case latencies [8]. This has considerable impact for reducing the overall power consumption and carbon footprint of data centres, which could consume 3 to 15% of global power by 2030 if power efficiency improvements are not made [11]. The implementation of optical switching in the data centre would be an important step towards achieving these power efficiency improvements.

However, a major challenge preventing the implementation of optically switched data centre networks is that the overall switching time of optical switches must occur on sub-nanosecond timescales to enable efficient handling of small packet dominated data centre traffic. This occurs because overall switching times above 10 ns in all-optical switches results in poor network utilisation [8], which is defined as the percentage of time that the network is used to send data, as opposed to time spent re-configuring the network. This acts to greatly reduce the performance, power consumption, cost and latency benefits resulting from using optical switching.

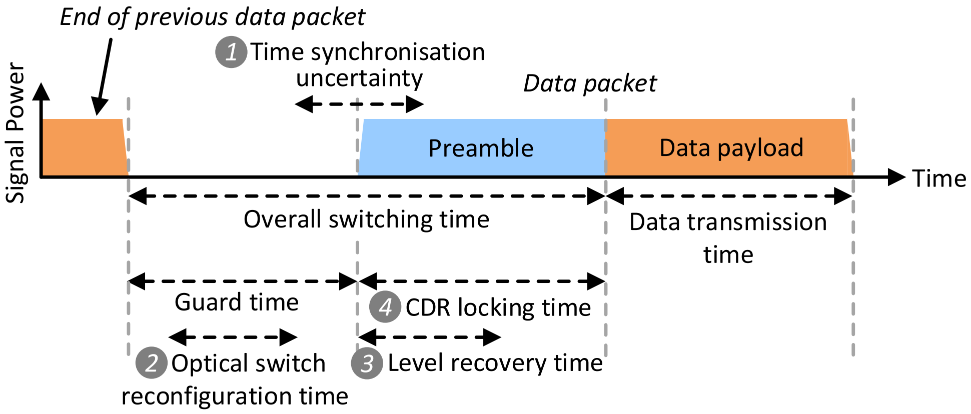

Fig. 1.2: The four components of overall switching time for packets arriving at a receiver connected to an optical switch. The overall optical switching time acts as overhead on the optical switch since data cannot be transmitted or received during this period. Note that the clock and data recovery (CDR) locking time and the level recovery time occur concurrently from the beginning of packet reception, and the required guard time between packets is the sum of the optical switch reconfiguration time and the time synchronisation uncertainty. A preamble is transmitted before the data payload to ensure that CDR locking and level recovery have completed before data reception.

As illustrated in Figure 1.2, the overall switching time of all-optical switches consists of four components, all of which must be minimised to sub-nanosecond to achieve high-performance operation in data centre optical switches:

① **Time synchronisation uncertainty:** the accuracy to which end-points connected to the optical switch are time synchronised.

② **Optical switch reconfiguration time:** the time taken for the optical elements in the fabric of the optical switch to reconfigure to create a new optical path.

③ **Level recovery time:** the time taken for receivers to recover the correct data sampling level after a data packet starts being received.

④ **CDR locking time:** the time taken for receivers to recover the correct sampling clock frequency and phase after a data packet starts being received.

Of these four components, only practically implementable sub-nanosecond CDR locking time has not been demonstrated. The previous state-of-the-art practical research prototype, based on digital phase interpolator CDR technology, achieved approximately only 5.8 ns CDR locking time at 56 Gb/s (325 symbols, equivalent to 12.7 ns at 25.6 Gb/s) [12]. In contrast, sub-nanosecond optical switching time has been demonstrated using fast-switching disaggregated wavelength tunable lasers [8, 13], SOAs [14] and Mach-Zehnder interferometers (MZIs) [15]; sub-nanosecond level recovery has been shown using caching of equaliser coefficients [8, 13] and sub-nanosecond clock synchronisation of 1000-node networks has been demonstrated through picosecond precision synchronisation techniques such as White Rabbit [16]. CDR locking time is therefore the key impediment preventing sub-nanosecond overall switching time, and is therefore the key impediment preventing the implementation of high-performance, practical data centre all-optical switching.

1.1.1 Clock and Data Recovery (CDR) Locking Time

The generalised concept of CDR and CDR locking time will now be explained. In any communications link where serial data is transmitted from one node to another without a separate, synchronously transmitted clock, such as in the generalised transmission link shown in Figure 1.3, the embedded, initially unknown clock frequency and clock phase must be extracted by the receiver using a CDR circuit, so that the data will be sampled at the correct frequency and the correct time position, as shown in Figure 1.4. This results in an initial period of incorrectly sampled data before the correct clock is extracted, with a high associated bit error rate (BER). The time between first reception of data and the first correct sampling of data is the CDR locking time as illustrated in Figure 1.2. After this period, the CDR circuit is then referred to as ‘locked’ since it will follow changes in frequency and phase of the embedded clock. The extracted clock is also typically used to supply a clock for circuitry that follows the data sampler, such as a deserialiser or data buffer.

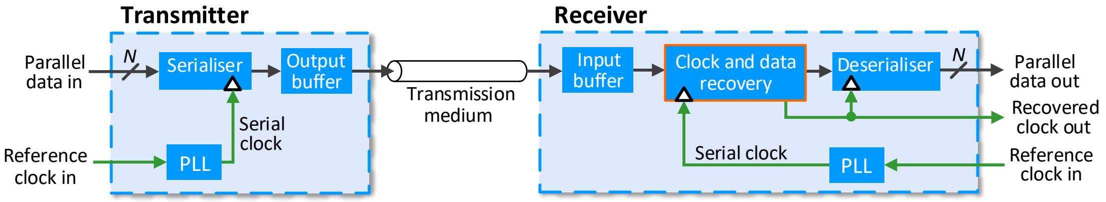

Fig. 1.3: Clock recovery in a generalised serial transmission link. The transmitted data contains an embedded clock, that must be extracted by the receiver using a clock and data recovery circuit (orange outline).

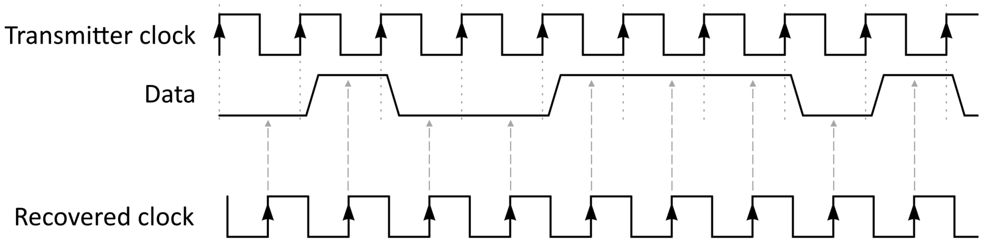

Fig. 1.4: Recovery of the embedded transmitter clock from received data. The recovered clock is phase and frequency aligned to sample the incoming data at the optimal sampling position halfway between transitions.

1.1.2 Impact of CDR Locking Time on Optical Switch Performance

In optical switches, the transmission of each data packet requires the creation of a momentary link between a transmitter-receiver end-point pair. The clock embedded in the data sent through each momentary link has a unique frequency and phase since the data originates from a new transmitter, incurring a CDR locking time period upon each switching event. CDR locking time results in a loss of network utilisation for optical switches (defined as the percentage of time that the link is used to send data), because data cannot be received without errors during the CDR locking time period.

In contrast, in current data centre networks built using electronic switches, the ASIC electronic switches and servers are connected by point-to-point serial links. In these standard data centre point-to-point links, a CDR locking time period is only incurred when each link is first brought online, since idle frames are transmitted continuously between data packets to ensure that the receiver is always able to follow any shifts in phase and frequency between the transmitter and receiver clocks. The long CDR locking time of current commercial CDR circuits therefore does not impact network utilisation in these networks[†].

As illustrated in Figure 1.5, with minimum size data packets, even if reconfiguring the optical switch takes only 1 ns, long locking time leads to low network utilisation. The network utilisation decreases as the CDR locking time increases because an increasing proportion of time is used for CDR locking rather than data packet reception. For example, assuming a 25.6 Gb/s transceiver, sending a minimum size 64-byte Ethernet packet only takes 20 ns [10]. Using a commercially-available CDR with 100 ns locking time, the overall network utilisation if 64-byte packets are received continuously would be less than 17%.

To understand the impact of CDR locking time on network utilisation for a practical optically-switched cloud data centre network, the distribution of packet sizes in the data centre traffic pattern that the optical switch would handle must be considered. Figure 1.6a shows a traffic distribution collected from a cloud data centre service, which is small packet dominated. Over 34% of the packets are less than 128 bytes, while 98% of the packets are equal or smaller than 576 bytes [19, 1]. This is consistent with a similar study from Facebook, in which 91% of the packets generated by their in-memory distributed cache service were 576 bytes or less [20].

[†] However, the continuous transmission of idle frames does have significant power consumption implications, since data centre links are typically heavily underutilised. For example, Facebook have reported that 99% of their data centre links are utilised under 10% of the time [17]. Not transmitting idle frames between data packets, such as in Energy-Efficient Ethernet [18], improves link power consumption but incurs a large latency penalty from long CDR locking time when the link is reactivated. Thus, even in electronically-switched data centre networks, though they are not the focus of this thesis, reduction of CDR locking time to nanoseconds would be beneficial [12].

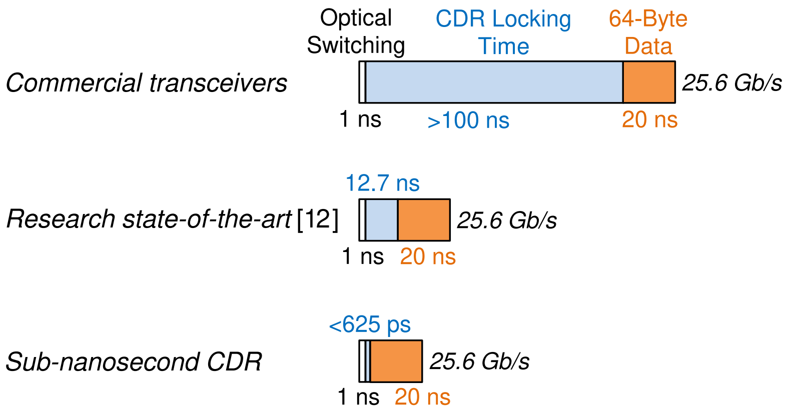

Fig. 1.5: The proportion of receiver time used by different CDR techniques when receiving minimum size (64-Byte) data packets. A large proportion of receiver time is spent handling minimum size Ethernet packets with current commercial and state-of-the-art research CDR approaches, reducing optical switch network utilisation. Sub-nanosecond CDR would allow this limitation to be overcome. Figure adapted from [1].

To evaluate the impact of the traffic pattern shown in Figure 1.6a, an event-based simulator was developed in-house at Microsoft Research Cambridge, which was cross-validated against real data-centre networks. The simulator modelled a network consisting of nodes interconnected by an optical switch. The line-rate of each node to/from the optical switch was 4×25 Gb/s. A synthetic workload was generated by randomly selecting the payload size for each packet from the distribution shown in Figure 1.6a and by selecting sources and destinations with a uniform random distribution. As soon as a node finished sending a packet, a new packet was generated and a new destination was selected. If a source-destination path was not already set up before the packet payload could be received, the optical switch needed to be reconfigured (leading to *optical switch reconfiguration time*) and the receiver CDR needed to lock onto the new incoming signal (*CDR locking time*). The optical switch reconfiguration time was set to 1 ns, based on recent advances in optical switching devices [15, 14, 13].

Figure 1.6b shows the impact of CDR locking time on optical switch network utilisation using the packet distribution in Figure 1.6a, generated using the event-based simulator by varying the CDR locking time. Reducing the CDR locking time from the state-of-the-art CDR locking time for phase interpolator CDRs of 325 symbols (12.7 ns at 25.6 Gb/s) [12] to sub-nanosecond would result in a 1.54× improvement in network utilisation. This improvement in network utilisation motivates the development of sub-nanosecond CDR, which will be the subject of this thesis.

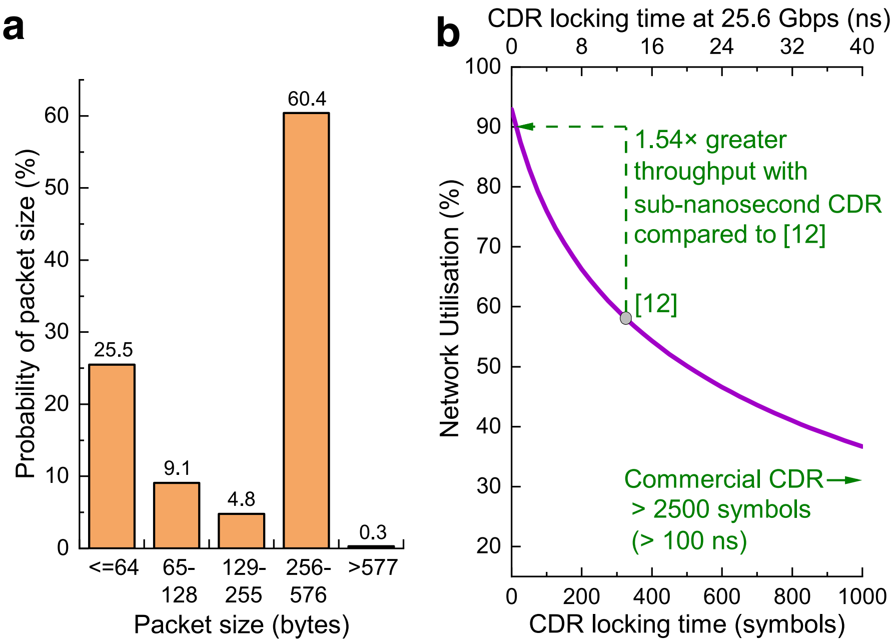

Fig. 1.6: Impact of CDR locking time on optical switch performance when handling real data centre traffic. **a**, Small packet dominated distribution of packet size in a measured production cloud data-centre traffic pattern. **b**, Gain in data-centre network utilisation from using sub-nanosecond CDR versus long CDR locking times when handling the traffic pattern shown in **a** (assuming 1 ns optical switch reconfiguration time). Figure adapted from [1].

1.2 Tools used in the Thesis

Examples and results from the analytical modelling described in this thesis were generated numerically using MATLAB and plotted using MATLAB and OriginLab. All experimental results were plotted using OriginLab. Microsoft Visio and Inkscape were used to create all illustrations. The thesis was written in and compiled with LaTeX.

1.3 Chapter Overview

The remaining Chapters in this thesis will cover the following topics:

Chapter 2 introduces the data centre environment, exploring electronic switching and optical transmission, as well as summarising research into data centre optical switching, establishing background for the remainder of the thesis.

Chapter 3 explores and evaluates existing methods of burst-mode CDR in the context of data centre optical switching, and explores approaches to synchronisation. It hypothesises that the need to lock to any random clock phase offset is limiting the achievable minimum clock recovery time for CDR circuits due to CDR metastability. Finally, *clock synchronisation assisted clock and data recovery (clock synchronisation assisted clock and data recovery (CSA-CDR))* is proposed, which could significantly lower CDR locking times by clock frequency and phase synchronising end-points connected through an optical switch, simplifying the CDR locking process.

Chapters 4 and *5* explore single calibration clock synchronisation assisted data recovery, where the transmitter to receiver clock phase offsets through an optical switch are calibrated once and are thereafter allowed to drift with changing data centre temperature. *Chapter 4* explores a first variation of this approach where the receiver CDR circuits are only used to measure the initial clock phase offsets, and are not used to track packet clock phase afterwards. A theoretical model of this variation is established, which explores the impact of changing data centre temperature on bit error probability at different data centre distance scales. *Chapter 5* explores a second variation of this approach, where the receiver CDR circuits do track packet clock phase after the initial calibration of clock phase offsets. The analytical model established in *Chapter 4* is extended to include the packet clock phase tracking behaviour of the CDR circuits, and to include the impact of high and low frequency contributions to jitter.

Chapter 6 explores ‘clock phase caching’, an approach to CSA-CDR where the CDR clock phase values are updated at regular intervals to compensate for fibre time-of-flight shift as the data centre temperature changes. The approach is first analytically modelled by extending the model in *Chapters 4* and *5*. Under 625 ps CDR locking time and error-free operation is then experimentally demonstrated in a real-time field programmable gate array (FPGA) based proof-of-concept 25.6 Gb/s non return to zero on off keying (NRZ-OOK) experimental testbed. The long-term reliability and the resilience of the approach to rapid temperature is demonstrated. The physical limits of the approach are explored under different emulated rates-of-change of data centre temperature, and under different magnitudes of clock jitter. Finally, the scalability of the ‘clock phase caching’ approach is explored.

Chapter 7 evaluates the benefits of using low thermal sensitivity hollow core fibre (HCF) with single calibration clock synchronisation CDR. Firstly, the state-of-the-art in HCF design, the guiding principle of HCF and the theory underpinning the 20× lower thermal coefficient of delay of HCF versus single-mode fibre (SMF-28) is described. The modelling in *Chapters 4* and *5* is then extended to consider the impact of the low thermal sensitivity of HCF. Experimentation to demonstrate the technique for a 2-to-1 optical switching system is then performed. Error-free operation with HCF over a 2 °C temperature range over 2 km of HCF and error-free 2-to-1 optical switching with under 625 ps CDR locking time over a 4 °C temperature range over 1 km of HCF are shown, with only an initial single calibration of clock phase. It is inferred from this result that a cluster spanning 100 m could operate error-free over a 20 °C temperature range, with no or an extremely small rate of clock phase updates. The potential impact of using HCF on the worst-case overhead of clock phase caching is also estimated.

Chapter 8 concludes the thesis, explores the potential impact of clock phase assisted clock and data recovery for optical communications and for other fields, and details possible future work that could be performed to build on the work shown in this thesis.

Appendix A provides mathematical derivations supporting the analytical modelling performed elsewhere within the thesis.

Appendix B provides an overview of the FPGA hardware constructed to implement and experimentally demonstrate CSA-CDR.

Appendix C provides photographs of the front and rear of the experimental setup used in Chapters 6 and 7 to demonstrate CSA-CDR.

1.4 Key Contributions

The key contributions of this thesis are:

- The thesis observes that a key factor limiting the performance of all optical switches is CDR locking time, and that it must be reduced to sub-nanosecond to maximise the performance of all-optical data centre switches.

- Existing methods of clock recovery are reviewed in a data centre context, and this thesis observes that no current method of CDR is viable for achieving sub-nanosecond CDR locking time. Synchronisation methods in optical time division multiplexing (OTDM), telecommunication networks and particle accelerators are reviewed on this basis. This thesis demonstrates *clock synchronisation assisted clock and data recovery (CSA-CDR)*, which significantly lowers CDR locking times by establishing frequency and phase synchronisation of transmitter to receiver pairs connected through an optical switch, simplifying the CDR locking process.

- An observation is made that in a real data centre environment slow changes in data centre temperature result in changes in optical fibre time-of-flight, which cause the correct clock phase values to shift. An analytical model is derived that models the effect on bit error probability from the shift in clock phase resulting from the changing temperature within the data centre environment, in the presence of optical noise and sampling clock jitter signal impairments.

- For SMF-28 fibre interconnection, the change in clock phase is approximately 40 ps/(km·°C) [21]. In the worst case, this is sufficient to cause almost half a bit period of shift at 25.6 Gb/s for a within-rack distance scale of 7 m across full data centre temperature temperature ranges of 40 °C, with proportionally larger shifts possible for longer distance scales. Thus, updating of the phase values is required on all data centre distance scales if SMF-28 transmission is used.

- An approach to CSA-CDR is proposed where the clock phase values are regularly updated, called *clock phase caching*. The bit error probability of a clock phase cached receiver, and the transmission overhead resulting from the technique, is analytically modelled, which suggested that the approach is viable.

- Clock phase caching is demonstrated in a point-to-point optical system and a 2-to-1 optically switched system with 2 km clock and 2 km data SMF-28. In both cases, it achieves CDR lock in under 16 symbols (625 ps at 25.6 Gb/s), an over 20× improvement on the previous state-of-the-art for phase interpolator CDRs of 325 symbols (equivalent to 12.7 ns at 25.6 Gb/s) [12], which is a highly commercialised, highly stable and practical CDR technology. This improvement in CDR locking time results in a 1.54× improvement in optical switch network utilisation versus the previous state-of-the-art CDR locking time when handing real data centre traffic with a 1 ns guard band between successive packets.

- The rate of phase updates required for clock phase caching as a function of rate-of-temperature change is experimentally evaluated. The tolerance of ‘clock phase caching’ to both random and deterministic clock jitter applied to the central synchronous clock source is evaluated.

- An analytical estimate of the network utilisation overhead resulting from performing the clock phase updates in clock phase caching is calculated, based on the rate-of-change of temperature, symbol rate and fibre thermal sensitivity. For a real production data centre thermal environment, the overhead is found to be only 2.2% for 10,000 end-points (servers or electronic switches) connected to an optical switch. This is sufficient to support 640,000 servers if 10,000 top-of-rack (ToR) switches connected to 64 servers each are connected to a single 10,000 port optical switch.

- A complementary approach to clock phase caching is proposed where the SMF-28 typically used in data centres is replaced with HCF, which has a 20× smaller thermal coefficient of delay than SMF-28 [21].

- The tolerance to temperature change of clock recovery operating in a point-to-point optical system with 2 km HCF and a 2-to-1 optically switched system with 1 km HCF is evaluated to be 2 °C and 4 °C respectively. Based on the analytical model, the tolerance to temperature change of a system interconnected by 100 m fibre would be 20 °C, sufficient to interconnect a single cluster of any number of servers without requiring clock phase updates over the entire industrially recommended data centre temperature range [22].

- The possible implications for overhead (and therefore scalability) from using HCF in a system using clock phase caching are evaluated.

- Future applications and extensions of clock synchronisation assisted CDR, both within the data centre, and in other applications, such as in the synchronisation of receivers in passive optical networks (PONs), are proposed.

1.5 List of Publications

A subset of the work presented in this thesis was first published in the following academic publications, in addition to in a patent:

Journal papers

1. **Kari A. Clark**, Hitesh Ballani, Polina Bayvel, Daniel Cletheroe, Thomas Gerard, István Haller, Krzysztof Jozwik, Kai Shi, Benn Thomsen, Hugh Williams, Georgios Zervas, Paolo Costa and Zhixin Liu. Synchronous Sub-Nanosecond Clock and Data Recovery for Optically-Switched Data Centres using Clock Phase Caching. *Nature Electronics* **3**, 426–433 (June 2020).

2. *(Invited, Highly-Scored)* **Kari A. Clark**, Yong Chen, Eric R. Numkam Fokoua, Tom Bradley, Francesco Poletti, David J. Richardson, Polina Bayvel, Radan Slavík and Zhixin Liu. Low Thermal Sensitivity Hollow Core Fiber for Optically-Switched Data Centers. *Journal of Lightwave Technology* **38**, 2703–2709 (May 2020).

Conference papers

1. *(Post-Deadline)* **Kari Clark**, Hitesh Ballani, Polina Bayvel, Daniel Cletheroe, Thomas Gerard, István Haller, Krzysztof Jozwik, Kai Shi, Benn Thomsen, Philip Watts, Hugh Williams, Georgios Zervas, Paolo Costa and Zhixin Liu. Sub-nanosecond clock and data recovery in an optically-switched data centre network. In *2018 European Conference on Optical Communication (ECOC 2018)* (Rome, Italy, September 2018).

2. *(Top-Scored)* **Kari A. Clark**, Yong Chen, Eric R. Numkam Fokoua, Tom Bradley, Francesco Poletti, David J. Richardson, Polina Bayvel, Radan Slavík and Zhixin Liu. Low Thermal Sensitivity Hollow Core Fibre for Optically-Switched Data Centre Applications. In *2019 European Conference on Optical Communication (ECOC 2019)* (Dublin, Ireland, September 2019).

Patent

1. Hitesh Ballani, Paolo Costa, Hugh David Paul Williams, István Haller, Krzysztof Jozwik, Benn Charles Thomsen, **Kari Aaron Clark**, Adam Christopher Funnell, Philip Michael Watts, Kai Shi and Thomas Michael Hoare Gerard. Phase caching for fast data recovery. US Patent 15857321, May 2019.

The following academic publications also include contributions that were made by the Author of this thesis during the completion of their PhD project, which are not presented within this thesis:

Journal papers

1. Ronit S. Sohanpal, **Kari Clark**, Benjamin J. Puttnam, Yoshinari Awaji, Naoya Wada, Polina Bayvel and Zhixin Liu. Clock and Data Recovery-Free Data Communications Enabled by Multi-Core Fiber With Low Thermal Sensitivity of Skew. *Journal of Lightwave Technology* **38**, 1636–1643 (April 2020).

2. Thomas Gerard, Hubert Dzieciol, Joshua Benjamin, **Kari Clark**, Hugh Williams, Benn Thomsen, Domaniç Lavery and Polina Bayvel. Packet Timescale Wavelength Switching Enabled by Regression Optimisation. *IEEE Photonics Technology Letters* **32**, 477–480 (April 2020).

3. Paris Andreades, **Kari Clark**, Philip M. Watts and Georgios Zervas. Experimental demonstration of an ultra-low latency control plane for optical packet switching in data center networks. *Optical Switching and Networking* **32**, 51–60 (April 2019).

4. Zhixin Liu, Boris Karanov, Lidia Galdino, John R. Hayes, Domaniç Lavery, **Kari Clark**, Kai Shi, Daniel J. Elson, Benn Charles Thomsen, Marco N. Petrovich, David J. Richardson, Francesco Poletti, Radan Slavík and Polina Bayvel. Nonlinearity-free coherent transmission in hollow-core antiresonant fiber. *Journal of Lightwave Technology* **37**, 909–916 (November 2018).

Chapter 2

The Data Centre Networking Environment

2.1 Introduction

This chapter provides background by exploring the data centre network environment. This background will consist of exploration of the following topics: the electronically-switched approach to networking in the data centre, priorities for and methods of optical transmission in the data centre, current and future approaches to optical transmission in the data centre, proposed approaches to optical switching in the data centre and a discussion of data centre temperature variation.

2.2 Electronically-Switched Data Centre Networks

In current data centres, electronic crossbar switches implemented on ASICs are interconnected to construct the network that allows servers to communicate with each other. Even with current state-of-the-art electronic switches, the total bandwidth of each electronic switch is small (currently up to 25.6 Tb/s, consisting of 64×400 Gb/s lanes [23]) in comparison to the sum of the link bandwidth of every server in a large data centre (for example, 100 Pb/s for a current large data centre with 4,000 racks of 64 servers [8]). This necessitates that thousands of electronic switches are interconnected to allow all servers to communicate with each other within a large data centre.

Although data switching within data centres is currently performed in the electronic domain, data transmission between servers and switches is predominantly performed in the optical domain. At all distance scales greater than those between servers and ToR switches within a rack, transmission over optical fibre is used. For relatively short distance scales of up to 100 m, for instance within a cluster or pod, transmission over multi-mode fibre (MMF) is typically used to minimise cost. For distances greater than this, for instance between the aggregation layer and the core, transmission over SMF-28 is used. NRZ-OOK or 4-level pulse amplitude modulation (PAM-4) direct detection modulation formats are currently used to avoid the additional power consumption and cost associated with coherent transmission and reception. Current approaches to optical transmission in data centres will be discussed in more detail later in this chapter.

Current data centre networks typically use folded-Clos topologies to interconnect servers using multiple layers of electronic switches [24]. The term folded-Clos is used because these topologies are logically equivalent to a unidirectional Clos network that is folded through the core switch layer. Figure 2.1a shows an example 3-layer folded-Clos topology, with 4 bidirectional ports per switch. Figure 2.1b shows the same topology re-arranged as a 5-stage unidirectional Clos network. Data centre networks are arranged in pods (or clusters) of servers, which contain edge (or ToR) switches to connect servers to the data centre network, as well as aggregation (or end-of-row) switches to interconnect the edge switches within each pod. Core switches then interconnect the aggregation switches. In topologies with over 3-layers, the Core layer is split into multiple layers.

The scalability of a folded-Clos topology is dependent on the number of layers in the switch hierarchy, L, and the number of ports per switch, k. The number of servers supported by a folded-Clos topology, as well as the number of servers and links required, can be calculated as follows [25]:

$$\text{Number of servers supported} = 2\left(\frac{k}{2}\right)^{L} \tag{2.1}$$

$$\text{Number of switches required} = (2L-1)\left(\frac{k}{2}\right)^{L-1} \tag{2.2}$$

$$\text{Number of transceivers required} = 4L\left(\frac{k}{2}\right)^{L} \tag{2.3}$$

Equations 2.1, 2.2 and 2.3 can also be combined to calculate the number of switches and transceivers required per transceiver:

$$\text{Number of switches required per server} = \frac{2L-1}{k} \tag{2.4}$$

$$\text{Number of transceivers required per server} = 2L \tag{2.5}$$

Consider a data centre network constructed of state-of-the-art 64×400 Gb/s electronic switches [23]. A 3-layer folded-Clos network constructed using these switches can support up to 65,536 servers with 400 Gb/s bandwidth per server, which requires using 5,120 switches and 393,216 transceivers. As this is not sufficient scalability to support a large data centre with 256,000 servers (4,000 racks of 64 servers), a further 4th layer is required [8]. A 4-layer folded-Clos topology with 64-port switches can support up to 2,097,152 servers, which requires using 163,840 switches and 16,777,216 transceivers. Although switch ports can be left unused to construct a subsection of the folded-Clos network (enabling a 256,000 server data centre to be interconnected), introducing the 4th layer still increases the number of transceivers required per server.

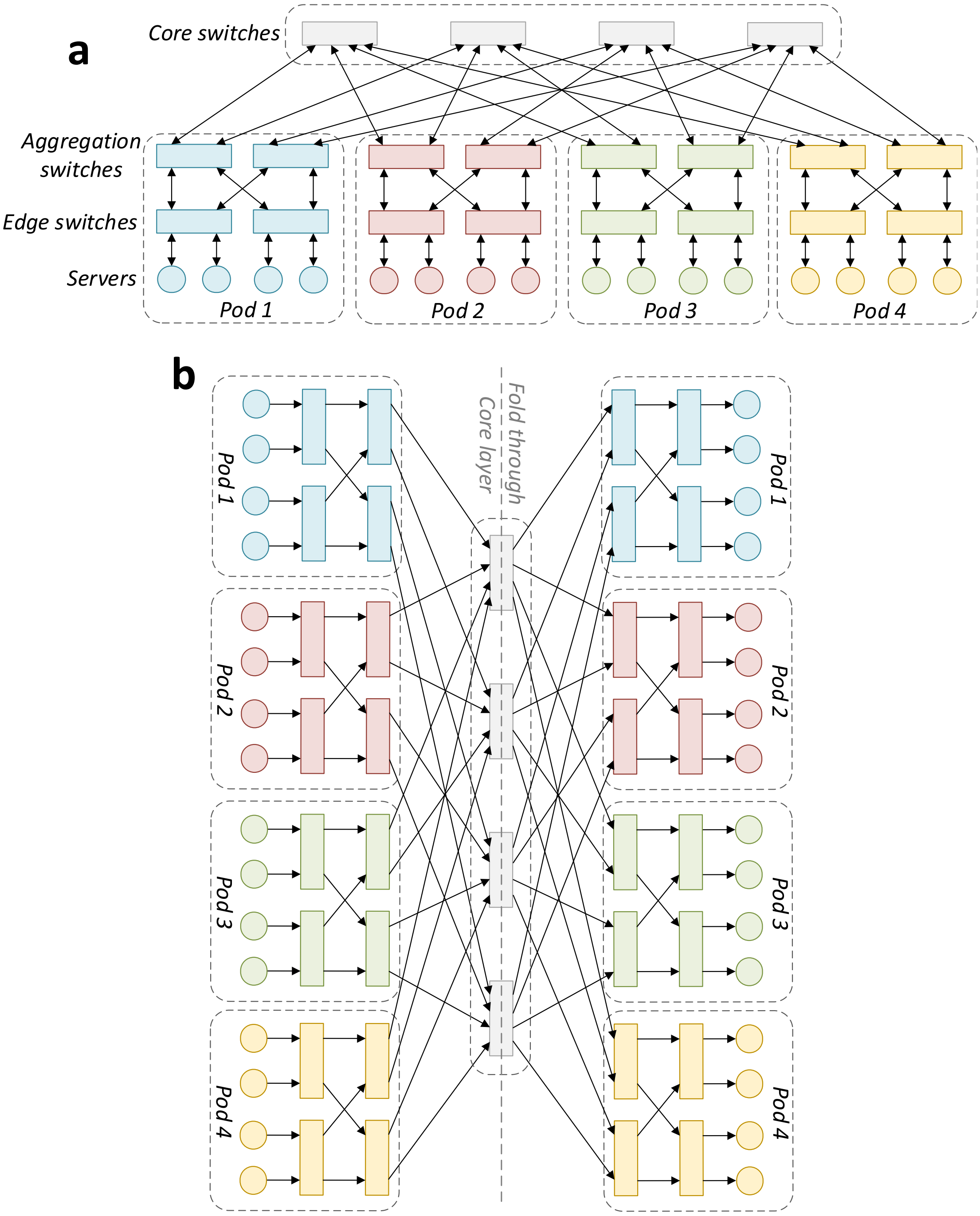

Fig. 2.1: Example folded-Clos data centre network topology used for electronically-switched data centre networking. **a**, 3-stage folded-Clos data centre network topology, with 4 bidirectional ports per switch. **b**, Equivalent unidirectional 5-stage Clos network, which is folded through the core switches to give the topology shown in **a**. Servers are illustrated with circles; switches are illustrated with rectangles.

The power consumption of such an ideal 4-layer network constructed of electronic switches would be prohibitive. For a 4-layer folded-Clos network constructed entirely using 400 Gb/s ports, the network elements required to interconnect each server has been estimated to contribute 194.8 W to overall data centre network power consumption [8]. For a large data centre with 256,000 servers, this would result in an overall data centre network power consumption of 50.3 MW. This exceeds the overall 32 MW allocation for an entire data centre [8]. Network latency and cost are also increased as a consequence of increasing the number of network layers [8].

To avoid the problems of high power consumption and cost, current data centres use over-subscription, where the overall network bandwidth of each higher-order layer is smaller than the previous lower-order layer, i.e. the network bandwidth of the core layer is smaller than the aggregation layer, which is in turn smaller than the edge layer. However, over-subscription results in resource fragmentation and an increase in worst-case latencies from tens of microseconds to hundreds of microseconds [26]. For the software-based workloads commonplace in current data centres, this has been tolerated as the latency resulting from software, which is on the order of milliseconds, has been much greater than the network latency [8]. Additionally, central processing unit (CPU)-based software workloads have been unable to saturate 100 Gb/s links between servers and their ToR switches [8].

However, emerging future workloads, such as deep neural network training, which are graphical processing unit (GPU)-based, will be limited by network latency (as they bypass the software stack) and are able to process 12.4 Tb/s of network traffic [27]. Furthermore, continued scaling of the switching bandwidth of each data centre electronic switch ASIC is ultimately limited by fundamental physical limitations on the power consumption and density of silicon transistors [6]. As a consequence of these fundamental physical limitations, the transistor power consumption benefit of reducing transistor node size has diminished for the most recent transistor node sizes, such as 7 nm, with less benefit gained with each successive reduction of node size following the ending of Dennardian scaling[†] [28]. This combination of demand from future high-bandwidth hardware-based workloads combined with electronically-switched network scalability limitations is driving research into optically-switched network alternatives that avoid the fundamental physical limitations of silicon transistors on electronic switch ASIC scalability.

[†] During the Dennardian scaling era, the increase in power density per node resulting from increasing transistor density ($\propto S^2$) and frequency ($\propto S$) was cancelled out by a simultaneous reduction in transistor gate capacitance ($\propto 1/S$) and transistor drain ($\propto 1/S^2$) to source voltage. In the current post-Dennardian scaling era, the transistor drain to source voltage has plateaued due to quantum tunnelling effects, resulting in an overall increase in power density per node ($\propto S^2$) [28].

2.3 Priorities and Properties of Data Centre Networks

To be viable for data centre networking, optical switching and transmission technologies must meet a set of priorities and properties that are specific to the data centre environment. These priorities and properties may be summarised as follows:

- **Low capital cost:** As data centres may contain hundreds of thousands of optical links, the capital cost of each link must be as small as possible, ideally under 1 $/(Gb/s), to minimise the overall capital cost of the data centre network. The number of switch elements also must be minimised.

- **Low power consumption:** As data centres may contain hundreds of thousands of optical links, the power consumption of each link and switch must be as small as possible, ideally under 1 pJ/bit [29], to minimise the overall power consumption of the data centre network.

- **Low latency:** Minimising latency through the data centre network is important as data centre application performance, particularly the performance of hardware workloads, is impacted by high network latency. Techniques that would increase latency such as soft-decision forward error correction (FEC) and digital signal processing (DSP) are minimised if possible due to their impact on latency.

- **Small bit error rates:** As soft-decision FEC techniques that would allow for compensation of small BERs, such as 10^{-3}, would introduce additional latency that would dominate over fibre transmission time, high signal quality transmission with small bit error rates of e.g. 10^{-12} to 10^{-9} are ideally targeted that can allow for either no FEC or hard-decision FEC [10].

- **Short transmission distance:** Transmission distances within a single data centre building or site are short, up to 7 m within a single rack, up to 100 m within a single cluster, up to 2 km for transmission within a single data centre building, and up to 10 km for transmission between nearby data centre buildings on a single site. The short transmission distance minimises the impact of optical impairments such as optical attenuation, wavelength dispersion and non-linearity.

- **Fault tolerance:** Hardware failures within the data centre must be tolerated without loss of service to end users. Failure of optical transceivers or switches could cause sections of the data centre to cease functioning. The probability of this occurring must be mitigated by using redundancy within the network.

2.4 Data Centre Optical Transmission

Current methods of data transmission in data centres are established by a combination of international standards and multi-source agreements, with the dominant standard being IEEE 802.3, the IEEE Standard for Ethernet [10]. There are four typical standardised transmission distance scales, illustrated in Figure 2.2 and summarised in Table 2.1:

- **Intra-rack / Edge:** The shortest distance within a hierarchical data centre network. A rack of up to ≈64 servers are interconnected by a single ToR switch. Standardised transmission lengths are up to 7 m, with transmission rates of 10 Gb/s non-return to zero (NRZ) signalling or 4×25 Gb/s using NRZ signalling, typically over twisted-pair copper cable to minimise cost [30, 31].

- **Intra-row / Intra-cluster / Pod:** Multiple ToR switches are interconnected by End-of-Row or spine switches, depending on network topology. A network of End-of-Row or spine switches forms a data centre cluster, interconnecting up to ≈10,000 servers in total. A single ToR switch will be connected to multiple End-of-Row or spine switches for redundancy. Standardised transmission lengths are up to 100 m, with transmission rates of 4×25 Gb/s using NRZ-OOK [32] or up to 8×50 Gb/s using PAM-4 modulation over MMF [33].

- **Intra-building / Core:** Multiple End-of-Row or spine switches interconnecting data centre clusters within a single data centre building, are in turn interconnected by core switches. Up to ≈250,000 servers may be interconnected by a network of core switches. A single End-of-Row or spine switch will be connected to multiple core switches for redundancy. Standardised transmission lengths are up to 2 km with transmission rates of 4×25 Gb/s using NRZ-OOK [34] or up to 8×50 Gb/s using PAM-4 modulation over SMF-28 [35].

- **Inter-building:** Multiple smaller data centres are interconnected to form a virtual data centre containing up to ≈1,000,000 interconnected servers. Older transmission standards are targeted at interconnecting data centre buildings located on the same building site, with standardised transmission lengths of up to 10 km, with transmission rates of 4×25 Gb/s using NRZ-OOK [31] or up to 8×50 Gb/s using PAM-4 modulation over SMF-28 [35]. Furthermore, due to the high cost of acquiring land for building large multi-building data centre sites, transmission between data centres that are located within the same metropolitan area, but not on the same building site, has also been standardised. Standardised transmission lengths are up to 80 km in this case, with transmission rates of up to 8×50 Gb/s using PAM-4 modulation [32].

Table 2.1: Summary of standardised data transmission scales in the data centre environment.

Data centre hierarchy level	Description	Distance scale	Typical number of servers	Current typical transmission medium	Current typical transmission rates
Intra-rack or Edge	Interconnection within a single data centre rack	Up to 7 m [30, 31]	Up to 64	Twisted-pair copper cable [30]	10 Gb/s NRZ [30] 4×25 Gb/s NRZ [31]
Intra-row, Intra-cluster or Intra-pod	Interconnection of top-of-rack switches within a data centre row or cluster	Up to 100 m [36, 33]	Up to 10,000	MMF [36, 33]	4×25 Gb/s NRZ-OOK [36] 8×50 Gb/s PAM-4 [33]
Intra-building or Core	Interconnection of end-of-row or spine switches	Up to 2 km [34, 35]	Up to 250,000	SMF-28 [34, 35]	4×25 Gb/s NRZ-OOK [34] 8×50 Gb/s PAM-4 [35]
Inter-building	Interconnection of data centre buildings	Up to 10 km [31, 35] Up to 80 km [32]	Up to 1,000,000	SMF-28 [31, 35, 32]	4×25 Gb/s NRZ-OOK [31] 8×50 Gb/s PAM-4 [35, 32]

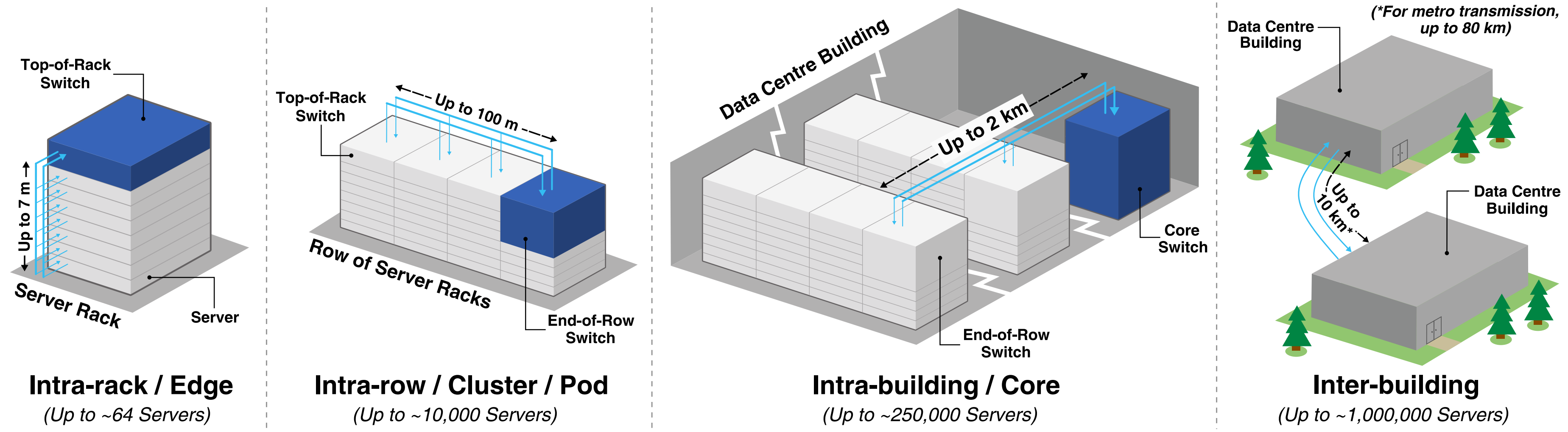

Fig. 2.2: Standardised data transmission scales in the data centre environment.

The current highest standardised data centre transmission rates are up to 400 Gb/s, reached using space division multiplexed 50 Gb/s (25 GBaud PAM-4) over MMF, or wavelength division multiplexed 50 Gb/s (25 GBaud PAM-4) over SMF-28. Current research in the intra-data centre transmission space continues to focus on increasing the bandwidth of each data centre transmission link, while simultaneously reducing overall link energy efficiency and component cost [29].

For transmission distances over short distances of up to 300 m, directly modulated 850 nm vertical-cavity surface-emitting lasers (VCSELs) are typically used for transmission over parallel MMFs with space division multiplexing (SDM). The bit rates of VCSELs modulated with NRZ have plateaued (the current record is 71 Gb/s [37]). Research to achieve per lane transmission rates of 100 Gb/s and above has focused on using higher-order modulation formats, such as PAM-4 and discrete multi-tone (DMT). Recent examples include >100 Gb/s error-free (BER $<10^{-12}$) transmission over 100 m MMF using PAM-4 [38] and 152 Gb/s transmission over 300 m MMF using DMT [39]. To extend overall link bandwidth to over 1 Tb/s while reducing the cost of the high number of parallel fibres required to do this, shortwave wavelength division multiplexing (SWDM) is a possible approach, with, for example, 4×100 Gb/s PAM-4 transmission demonstrated over a single 100 m MMF [40].

Over longer intra-data centre scales of up to 2 km, where SMF-28 is used as the transmission medium, 200 Gb/s per-wavelength transmission using PAM-4, 8-level pulse amplitude modulation (PAM-8) and DMT modulation formats has been demonstrated, though with considerable FEC requirements, as reaching per-wavelength transmission rates of 200 Gb/s pushes intensity modulation direct detection (IM-DD) transceiver components to their limits [41]. For future advancement in data rate per fibre beyond 1 Tb/s, Dense Wavelength Division Multiplexing (DWDM) implemented using parallel in-plane transmitters or optical frequency combs is one possible approach, and another alternate or complementary approach, if coherent transceiver complexity is sufficiently reduced, is to move to coherent transmission formats instead of the IM-DD approach that is currently used in intra-data centre transmission [42]. The potential for reduced transceiver cost by using silicon photonics technologies to improve integration of optical transceiver components has also attracted strong recent attention [43].

At the time at which the results in this thesis were obtained (2017 to early 2019), 50 Gb/s transceivers were not yet available commercially in Xilinx FPGAs. FPGA transceivers would be crucial for the demonstration of real-time clock phase synchronised optical switching at 50 Gb/s. As an consequence of this, this thesis will focus on minimising CDR locking time using 25 GBaud NRZ-OOK data transmission throughout, using Xilinx 25 GBaud NRZ FPGA transceivers [44] that were the fastest commercially available at the time of result acquisition.

2.5 Optically-Switched Data Centre Networks

This section will provide an overview of proposed approaches to implementing optical switching within the data centre networking environment. This discussion will be kept short and high-level, as achieving sub-nanosecond CDR is the core focus of this thesis, which is a crucial requirement for nanosecond speed all-optical switches irrespective of network topology and switching technology. More comprehensive details on these approaches can be found in the following review papers and theses: [29, 45, 46, 47].

In optical switching, momentary optical paths are created between servers or switches to transmit data directly between them, rather than through a network of interconnected electronic switches. Using this approach avoids the scalability concerns associated with electronic switches, allowing continued scaling of data centre bandwidth while minimising power consumption, latency and cost. Many different approaches to implementing optical switching in the data centre have been proposed. These approaches can be broadly grouped by the following characteristics:

- **Optical switch reconfiguration speed / Optical switch technology:** Different technologies used to implement optical switching have different optical switch reconfiguration times, from sub-nanosecond to millisecond.

- **Packet vs circuit switching:** In circuit switching, the paths through the optical switch are allocated, and the switch reconfigured, prior to transmission. In packet switching, data is routed through the optical switch on a per-packet basis.

Microelectro-mechanical system (MEMS) optical switches, with slow optical reconfiguration times on timescales of microseconds to milliseconds, can be used to perform slow circuit switching, as illustrated in Figure 2.3. The slow optical reconfiguration times of MEMS optical switches come from their principle of operation, in which micro to millimeter-scale mirrors are physically re-orientated to alter the path of light through switch. Hybrid data centre networks have been proposed that consist of slow-switching optical circuit switches based on MEMS technologies operating in parallel with a separate fast-switching electronically-switched network. Long flows, of lengths significantly longer than the overall optical switching time of MEMS optical switches, would be transmitted on the optically-switched network, while short flows would be transmitted on the electronically-switched network. Prominent examples of hybrid switched architectures include c-Through [48], Helios [49] and Mordia [50]. MEMS optical switches with up to 384×384 ports are available commercially (although at high capital cost), with a worst-case optical loss of 2.7 dB and a switching time of 25 ms [51].

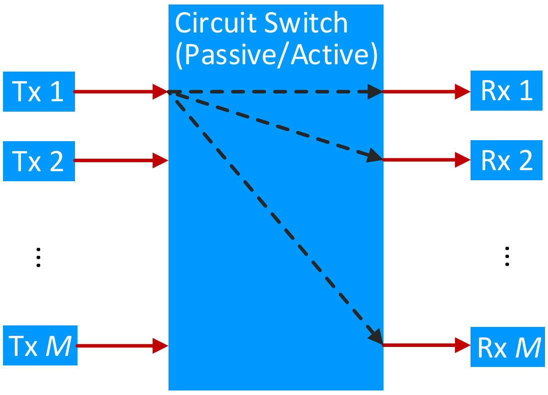

Fig. 2.3: Principle of optical circuit switching. Optical paths are formed through the switch core to communicate between transmitters and receivers connected to the optical switch. The switching can be performed using wavelength or space switching at the edge transmitters and/or receivers (with a passive core consisting of star couplers and/or arrayed waveguide grating routers (AWGRs)), within the core itself (an active core), or using a combination of edge and core switching. The optical circuit switching speed can be slow (microsecond to millisecond, using MEMS switches), or fast (nanosecond), using SOA, tuneable laser and/or MZI switching elements.

Due to the microsecond to millisecond optical reconfiguration time of MEMS-based optical switches, microsecond-long typical CDR locking times for commercial transceivers do not limit their performance, so these switches do not require sub-nanosecond CDR. However, there are challenges associated with the scheduling of flows between the two network technologies in hybrid optical/electronic networks, including the need for rapid flow identification (required to determine whether a flow should be handled by the electronic or the optical network), handling flow correlation (where different flows are interdependent), and handling traffic patterns with many bursty flows (which consist of many short flows of millisecond length) [52]. However, network control planes that can react on 100s of microsecond timescales [53] as well as static scheduling [54] are promising potential solutions.

Optical technologies that can be used to construct optical switches with nanosecond optical reconfiguration times have also been demonstrated: MZIs [15], SOAs [14] and tuneable lasers [8, 13]. By operating on nanosecond timescales, individual data packets can be transmitted through networks constructed using these components, avoiding the need for a separate electronically switched network, in-turn avoiding the associated flow scheduling challenges associated with hybrid networks. Transmission of individual packets through a nanosecond optical switch can be achieved using a packet-switched approach, circuit-switched approach or a hybrid that combines the two approaches.

In the packet-switched approach to nanosecond optical switching illustrated in Figure 2.4, a label containing the destination node is transmitted with the data packets. This label is extracted at intermediary nodes, which is read and then used to switch the data packet to an appropriate optical path. The data packets themselves are delayed while the routing decision is made, optically or electronically. Optical packet switching was widely researched in the 2000s. Example architectures that use this approach include the Data Vortex [55], SPINet [56] and LIONS [57]. A significant current limitation of this approach for data centre applications is the practicality of the packet delay method. In some packet switched architectures, such as in the Data Vortex [55] and SPINet [56], optical fibre delay lines are used to delay the optical data packets, due to the lack of practical alternatives for optical buffering at the time of publishing. These would require large quantities of optical fibre for constructing large scale optical switches, which would be both space and cost inefficient. Other architectures, such as LIONS [57], implement the delay electronically instead, which would introduce additional optical-electrical-optical (OEO) conversions, which would in turn increase both power consumption and cost, diminishing the benefits of optical switching. However, significant progress has been made in optical memory technology in the 2010s [58], which may in the future lead to new research into optical packet switching using such technologies.

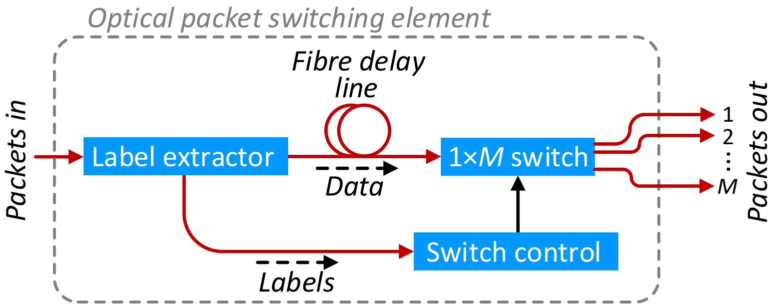

Fig. 2.4: Principle of optical packet switching. In optical packet switching, a label indicating the destination is stripped from incoming packets, which are then used to determine how to configure the switch. The data payload is delayed (optically or electronically) while the decision and configuration occurs. The figure shows a single optical packet switching element; many such elements are used to construct an optical packet switch. The optical switching speed must be fast (nanosecond), using SOAs, tuneable lasers and/or MZIs as switching elements.

In the circuit-switched approach to nanosecond optical switching illustrated in Figure 2.3, momentary optical paths are formed through an optical switch through which data is transmitted through. To perform this, passive components, typically AWGRs and star couplers, are used in combination with active components, wavelength tuneable lasers and SOAs/MZIs, to perform wavelength and space switching respectively. Unlike in optical packet switching, no buffering, optical or electronic, occurs within the optical switch core. Older examples of nanosecond circuit-switched architectures, such as OSMOSIS [59], used discrete component SOAs to perform both wavelength and space switching, which would not be power or space efficient for large scale optical switches.

Recent improvements of the worst-case all-to-all wavelength tuning times of tunable lasers using current waveform shaping [60], combined with optical integration with photonic integrated circuits (PICs) [61], have led to the development of PICs with integrated tunable lasers and SOAs that can perform cost and power efficient sub-nanosecond wavelength and space switching [13]. PULSE [62] and Sirius [8] are two optical switch architectures that use fast tuneable lasers and SOAs to perform optical circuit switching. A significant challenge for nanosecond optical circuit switching was the complexity of the optical switch scheduling, but recent research on distributed hardware schedulers [62], parallel modular schedulers [63] and coordination-free (cyclic) schedulers [8, 64] have provided promising solutions.

Three key general observations for this thesis are: *1)* all types of optical switches require synchronisation of all end-points connected to the switch to avoid packet collisions at the receivers [65], *2)* all optical switches require re-acquisition of the sampling clock following the establishment of a new optical path between end-points [65] and *3)* irrespective of whether the packet switching or circuit switching approach to nanosecond optical switching is used, sub-nanosecond CDR is crucial for achieving nanosecond optical switching. Sub-nanosecond CDR was achieved in Sirius [8] using the clock phase caching technique presented in this thesis.

As yet, major commercial data centre operators have not began constructing optically-switched data centre networks, or at least have not revealed public information indicating that they are doing so. However, pressure from bandwidth requirements of future hardware workloads, bandwidth limitations of future electronic switches, and potential power consumption and cost benefits of optical switching are driving major commercial data centre operators, such as Microsoft, to invest in the construction of prototype small-scale data centre optical switches that emulate the behaviour of a larger commercial optically-switched network [8]. Of the two approaches to data centre optical switching that are currently undergoing the most active research, the slow (hybrid, microsecond to millisecond optical reconfiguration time) and fast (all-optical, nanosecond optical reconfiguration time) approaches to optical circuit switching, hybrid approaches are likely to be implemented in data centre networks first, as they can be constructed by adding a second optically-switched network atop the existing electronically-switched network. All-optical switching, which would likely offer superior improvements in power consumption, cost and performance, but requires the complete replacement of the electronically-switched data centre network, is likely to be implemented in data centres at a later stage.

2.6 Comparison of Data Centre Network Approaches

Table 2.2 compares different approaches to constructing data centre networks.

Table 2.2: Comparison of data centre network approaches.

Data Centre Networking Architecture	Electronic Switching (Current Approach)	Hybrid Electronic Switching / Slow Optical Circuit Switching	Optical Packet Switching	Ultra-Fast Optical Circuit Switching
Description	Many small capacity electronic packet switches are interconnected by optical transmission links to construct network.	An electronically switched network is combined with a separate slowly switching optical circuit-switched network.	Optical packets containing a label and payload are routed by a series of optical packet switch elements.	Optical transmissions are sent through an ultra-fast configured optically-switched core.
Advantages	• Established technology. • High efficiency switching of small packets. • Only optical transmission, no additional complexity from optical switching. • Compatible with current networking software stack.	• Simpler integration with current electronically-switched networks. • No requirement for sub-nanosecond CDR. • Better compatibility with software stack than all-optical approaches.	• Simplicity of scheduling: performed at each node. • Security: data switched in the optical domain. • Power consumption improvements likely.	• Can be constructed with current optical technologies. • Potential for passive core: can just upgrade edges. • Known power consumption, cost and latency benefits. • Security: data switched in the optical domain.
Disadvantages	• Network scalability constrained by silicon transistor scalability • Over-subscription required to reduce network power consumption and cost to reasonable levels. • Poor latency: deep hierarchy of switches required.	• Scalability of fast-switching electronic network still constrained by silicon transistor scalability. • Scheduling: difficulty distinguishing between short and long flows. • High cost of MEMS switches.	• Requires optical delay lines (large, expensive), optical memory (not mature technology) and/or electronic memory (high power consumption). • Active core: cannot just upgrade edges.	• Complexity of scheduling 1000s of endpoints. • No large-scale (>1000 end-point) demonstrations with sub-nanosecond optical switching. • Large changes to software networking stack and associated software likely required.

2.7 Data Centre Temperature Variation

Data centre servers and networking equipment power consumption transfers heat to the data centre environment, which must be removed to prevent equipment failure. This is achieved by maintaining server and switch intake temperatures within tolerable limits using air conditioning. Changes in the data centre thermal environment affects synchronisation of servers and switches connected to an optical switch, as changing fibre temperature causes a proportional change in fibre time-of-flight. This effect is significant, since the thermal coefficient of delay of SMF-28 is ≈40 ps/(km·°C). 1 °C of temperature change across 1 km of SMF-28 therefore causes 40 ps time-of-flight change, which is of equal length to 1 symbol period at 25 GBaud. The material introduced in this section will be referred to in later Chapters in this thesis.

Figure 2.5 shows a typical topology for data centre cooling: hot aisle containment. In hot aisle containment, thermally controlled cool air located in the space between racks, called the cool aisle, is drawn through the air intake at the front of servers and switches contained in racks. The exhaust from the rear of the servers and switches is emitted into the hot aisle. The hot air from server and switch exhausts is then cooled to a set-point temperature by computer room air conditioner (CRAC) units, and once cooled, is emitted back into the cool aisle. To maximise energy efficiency, the hot and cool aisles are physically separated from each other.

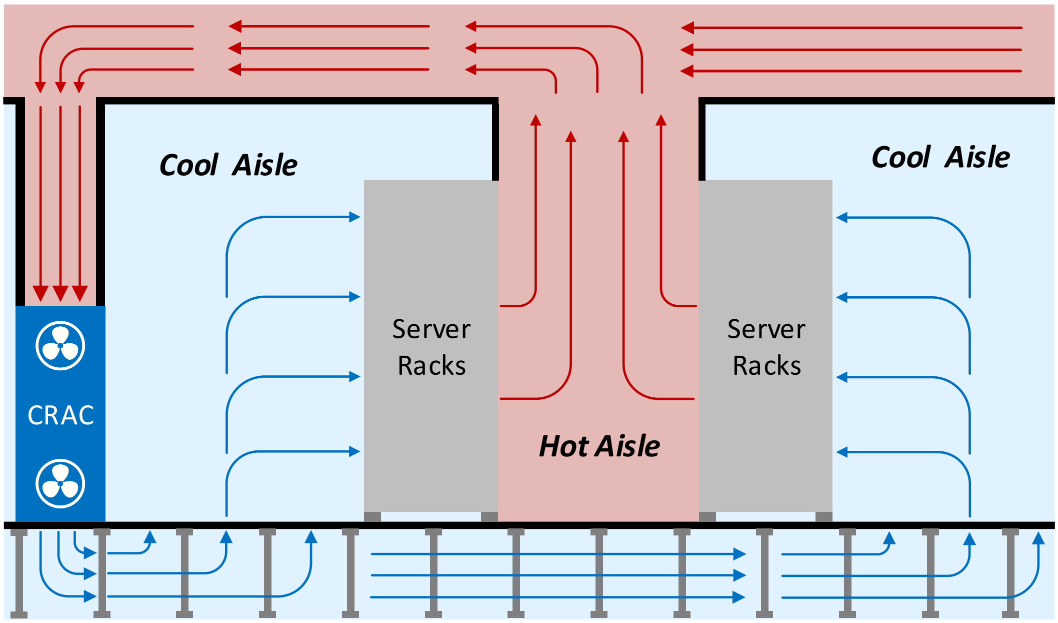

Fig. 2.5: Hot aisle containment: a typical modern data centre cooling topology. Hot and cold air are separated, to maximise cooling efficiency. The temperature of the air within the cool aisle is controlled by CRAC units. Server air intakes draw cool air from the cool aisle. The heated server exhaust is output into the hot aisle, and is then drawn back into the CRAC units to be cooled. CRAC; computer room air conditioner. Illustration adapted from [66].

Data centre design standards, such as TIA-942-B [67], include requirements for data centre temperature that are kept in parity with recommendations from professional air conditioning associations such as the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). Current data centre recommended server intake temperatures from ASHRAE are between 18 and 27 °C for optimal operation, with maximum allowable temperatures of 15 to 32 °C for operational equipment, and maximum allowable server intake temperatures of between 5 and 45 °C for server equipment that is switched off [22]. These larger ranges exist to maximise data centre cooling efficiency while still maintaining an acceptable rate of equipment hardware failure [22]. Fibre conduits carrying fibre to data centre core switches are typically located in the cool aisle space supplying cold air to data centre equipment, so that they are safely accessible by data centre operators. The clock and data fibre running to/from a data centre optical switch and top-of-rack switches could therefore experience worst case temperature variation between 5 and 45 °C.

Server exhaust temperatures are not controlled as they are dependent on data centre equipment workload. However, surveys performed of data centre networking equipment have found that some small single unit ToR switches can potentially have exhaust temperatures of 75 °C, which can lead to localised hot-spots at the rear of the server rack [68]. Minimum temperatures would be equal to minimum intake temperature (as data centre equipment can only heat air passing through it). Fibre cabling routed within the hot aisle side of the rack could therefore in the worst case be subjected to temperatures ranging from 5 to 75 °C. Optical fibre between a clock phase synchronised rack-scale optical switch and connected servers could potentially experience this worst-case temperature range.

An additional consideration is the maximum rate-of-change of temperature within a data centre. ASHRAE recommend that air intake temperatures in data centres do not change by any more than 5 °C per hour in a data centre that includes tape drives (which are vulnerable to rapid changes in temperature), and not by any more than 20 °C per hour (a maximum of 5 °C in any 5-minute period) in data centres that only include hard disc drives [22]. There are no recommendations for maximum rate-of-change of equipment exhaust temperatures.

Chapter 3

Burst-Mode Clock and Data Recovery (CDR), Clock Synchronisation Approaches and Proposed Approach

This chapter first explores current approaches for constructing burst-mode CDR circuits, aimed at minimising CDR locking time in data centre optical switches, and then evaluates their suitability for sub-nanosecond optical switching in the data centre environment. The characteristics and overall shortcomings of current burst-mode CDR approaches will then be summarised. Methods of synchronisation as used in optical networks and other related fields are then explored. Lastly, a new approach to burst-mode CDR is proposed, based on clock synchronisation assistance.

3.1 Requirements for Burst-Mode CDR Circuits

For practical usage in intra-data centre all-optical switches, CDR circuits would ideally need to satisfy the following properties [69]:

- Sub-nanosecond CDR locking time, to maximise optical switch network utilisation.
- Low power consumption and small silicon area usage to minimise overall transceiver (and by extension data centre) power consumption.
- Consistent performance in the presence of process voltage temperature (PVT) variation.
- Operation at current data centre transmission rates of 25 Gb/s and above.
- Ability to operate at slower data rates such as 10 Gb/s in addition to higher rates, for fault tolerance purposes.
- The ability to tolerate frequency offsets of up to 200 ppm from frequency variation between different source oscillators.
- The ability to place multiple CDR circuits next to each other, which correspond to multiplexed transceivers, without risk of crosstalk between the CDRs.

3.2 Burst-Mode CDR Approaches

3.2.1 Gated-Voltage Controlled Oscillator

Gated voltage controlled oscillator (GVCO) CDRs, illustrated in Figure 3.1, can have sub-nanosecond CDR locking times, but have significant limitations rendering them impractical for data centre applications. GVCO CDRs typically contain two voltage controlled oscillators (VCOs). One oscillator is built within a phase locked loop (PLL) circuit to lock to an external reference source to generate a clock signal with a frequency matched to that of the incoming data. This provides a frequency reference for a second oscillator, the GVCO, which is gated by transitions in incoming data. After being gated, the GVCO outputs a clock that is used to sample incoming data until the next transition in the data arrives [70].

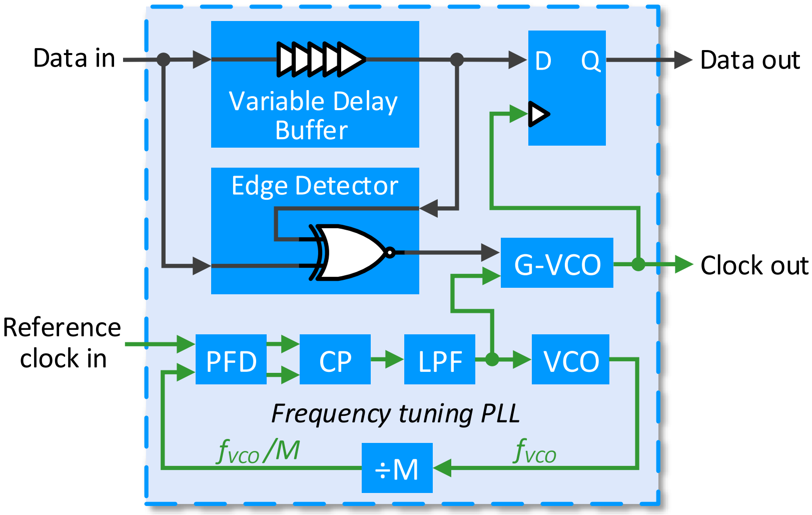

Fig. 3.1: GVCO CDR architecture. A frequency tracking PLL locks to a reference clock close to, or close to a division of, the embedded clock frequency of the incoming data, to provide a reference for the GVCO. The GVCO is then phase realigned by edges in the incoming data signal. G-VCO, gated voltage controlled oscillator; PFD, Phase frequency detector; CP, charge pump; LPF, low pass filter; VCO, voltage controlled oscillator. Figure adapted from [71].

In [72], a CDR is demonstrated that operates at 33.8 Gb/s that achieves CDR lock in 0.2 ns. However, no BER performance nor justification of the 0.2 ns locking time is provided and the validity of the result is therefore unclear. In [73], a CDR is demonstrated that operates with a BER of under 10^{-12} for a pseudo random binary sequence (PRBS)-7 sequence at 10.3 Gb/s that achieves CDR lock in a single symbol period. In [74], a CDR is demonstrated that operates with a BER of under 10^{-12} for a PRBS-31 sequence at 10 Gb/s that achieves CDR lock in 0.5 ns. In [75], an injection-locked CDR is demonstrated using two cascaded gated injection-locked VCOs that achieves CDR lock in one bit at 20 Gb/s. Although a BER of under 10^{-12} is demonstrated for a 2^7-1 PRBS sequence, the demonstrated BER for a 2^{31}-1 PRBS sequence is in contrast only 10^{-9}, indicating a possible BER penalty from long lengths of consecutive identical digits (CID).

A general limitation of GVCO CDRs for data centre all-optical switching is that frequency mismatch between the incoming data and the natural frequency of the gated oscillator, or natural frequency mismatch between the GVCO and the reference VCO, causes reduced tolerance to long lengths of CIDs (which may be up to 65-bits long in 64B/66B-encoded data centre links). This occurs because any accumulated phase mismatch between the embedded clock in the incoming data and the generated clock is corrected only when a transition occurs. PVT variation causes frequency mismatch between the oscillators, reduced tolerance to CIDs and therefore causes increased BER [69, 75]. Removing the frequency tuning PLL and controlling the gated-oscillator natural frequency from a digital to analog converter (DAC) may help address this disadvantage [73] but at the cost of introducing additional power consumption and circuit complexity.

In addition to this limitation, each CDR must have a dedicated LC-tank oscillator to serve as the GVCO for high-performance operation at 25 Gb/s data rates, which results in increased CDR silicon area and power consumption over alternative CDR designs such as the digital phase interpolator CDR that does not require a dedicated oscillator for each lane [69]. Additionally, crosstalk between adjacent CDRs in different receivers may also occur as a result of using a dedicated oscillator per transceiver [69]. The data rate must also be matched to the natural frequency of the injection oscillator, limiting transceiver data rate flexibility, which is important in data centre networks [69]. GVCO CDRs also have poor jitter rejection characteristics, passing all jitter present in the input data signal to the output sampled data [69].

3.2.2 Oversampling

Oversampling-based CDRs, as shown in Figure 3.2, avoid the necessity to recover the transmitter clock from received data by sampling it multiple times per symbol period [71]. There are two methods of oversampling: oversampling in time, where the data is sampled N times per symbol period with a single data sampler, which is driven by a sampling clock that is N times greater frequency than the data symbol rate; and oversampling in space, where the data is again sampled N times per symbol period, but by using N data samplers that are triggered by N evenly-spaced clock phase shifted versions of a line-rate sampling clock [76]. In both types of oversampling CDR the optimum clock phase to sample the data is selected each symbol period to sample the data. The data must be sampled at at least doubled the data symbol rate.

Oversampling in time is considered impractical for 25 Gb/s and above data rates due to the high circuit complexity and high power consumption resulting from the very high electronic clock frequency required to drive the data sampler [69]. Although oversampling in space has been demonstrated at bit rates of 5 to 10 Gb/s with instantaneous CDR locking time for use in PON applications [76, 77], the output clock experiences large jumps in phase between successive symbol periods which causes the jitter properties of the recovered clock to be poor [69]. Additionally, when a small number of sampling phases is used, such as 2, the sampling phase may deviate significantly enough from the optimal sampling phase to result in a significant BER penalty (the impact of sampling clock phase offset on BER is explored in Chapter 4).

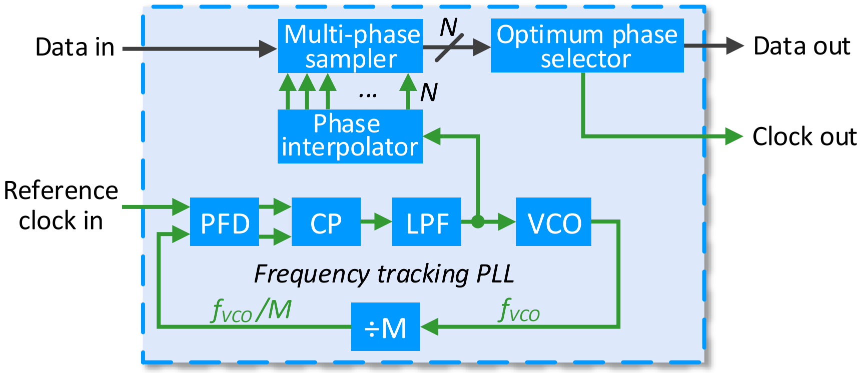

Fig. 3.2: Oversampling CDR architecture. A frequency tracking PLL locks to a reference clock close to, or close to a division of, the embedded clock frequency of the incoming data. In oversampling in space (as shown in this Figure), a phase interpolator then samples the incoming data at N equally-spaced phases of the incoming data. In oversampling in time, the data is sampled using a single data sampler with a sampling frequency N times greater than the data symbol rate. In both approaches, an optimum phase selector then selects the ideal phase with which to sample the data. PFD, phase frequency detector; CP, charge pump; LPF, low pass filter; VCO, voltage controlled oscillator. Figure adapted from [71].

3.2.3 Digital Phase Interpolator

Digital phase interpolator (PI) based CDR circuits have many advantages for continuous intra-data centre transmission, and are therefore widely used in commercial transceivers, such as in Xilinx UltraScale GTY FPGA transceivers [44]. Multiple CDRs can share the same reference clock (e.g. generated by a single LC-tank oscillator) [44], and can be entirely digital, minimising silicon area and power consumption [71]. They are also able to operate at any data rate up to the specification of the receiver [71].

The operational principle of a PI based CDR architecture, shown in Figure 3.3, is as follows: A PI based CDR takes a external reference clock that is closely matched in frequency to the incoming data signal (e.g. within 200 parts per million (ppm)). A phase interpolator (also known as a phase rotator) then generates N clock outputs that are phase shifted versions of the reference clock (e.g. 64 phases in steps of $\frac{1}{64}$ symbols). A phase selector selects a phase of the reference clock. The phase of the reference clock is compared to transitions present in the incoming data with a bang-bang phase detector (BB-PD). Every data transition, the BB-PD outputs a signal to indicate whether the phase-shifted reference clock was early (arrived before the data transition), or late (after the data transition). The digital controller counts early and late signals across a set of successive bit periods (e.g. 32, 64 or 128), and measures the difference between early clock transitions to late clock transitions. The reference clock phase supplied to the BB-PD is then shifted by a number of phase steps proportional to this difference, with the proportionality factor between these called the CDR proportional gain constant. Lastly, the data is sampled using a phase of the reference clock that is shifted $\frac{1}{2}$ symbols from clock phase supplied to the BB-PD.

However, standard commercialised digital phase interpolator based CDR circuits typically have worst-case CDR locking times of 100s of nanoseconds or longer. For example, for Xilinx 25 Gb/s UltraScale GTY Transceivers, typical locking times are stated as 50,000 symbols (2 μs at 25 Gb/s) and maximum locking times are stated as 2.3 million symbols with dynamic feedback equalisation disabled and 37 million symbols with dynamic feedback equalisation enabled (92 μs and 1.5 ms respectively at 25 Gb/s) [78]. These long CDR locking times arise from metastability of the BB-PD as illustrated in Figure 3.4, which occurs when the initial data sampling phase offset is close to $\frac{1}{2}$ a symbol from the correct sampling position [12]. Sampling clock jitter causes reduction of the expectation of the clock phase error from the ideal clock phase error output values of -1 or 1 (equivalent to the early or late signals discussed previously).

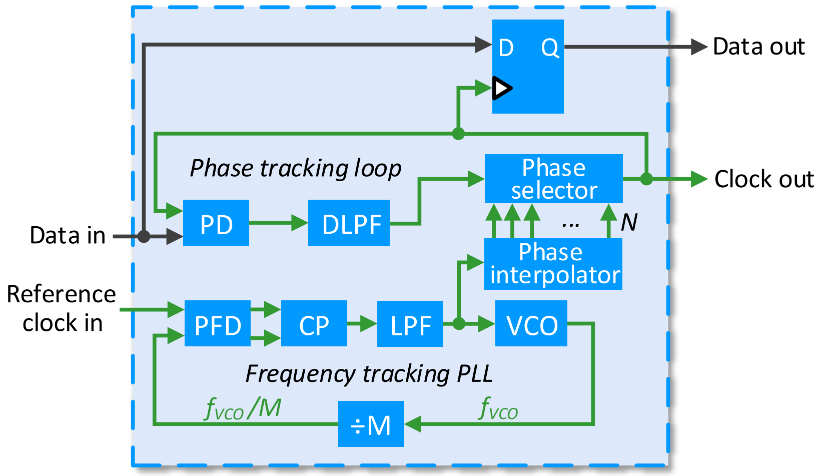

Fig. 3.3: PI CDR architecture. A frequency tracking PLL locks to a reference clock close to, or close to a division of, the embedded clock frequency of the incoming data. Multiple phases of this reference clock are then generated using a phase interpolator. A phase tracking loop containing a phase detector (typically a BB-PD) and a phase selector is used to track the optimal phase with which to sample the incoming data. PD, phase detector; DLPF, digital low pass filter; PFD, phase frequency detector; CP, charge pump; LPF, low pass filter; VCO, voltage controlled oscillator. Figure adapted from [71].

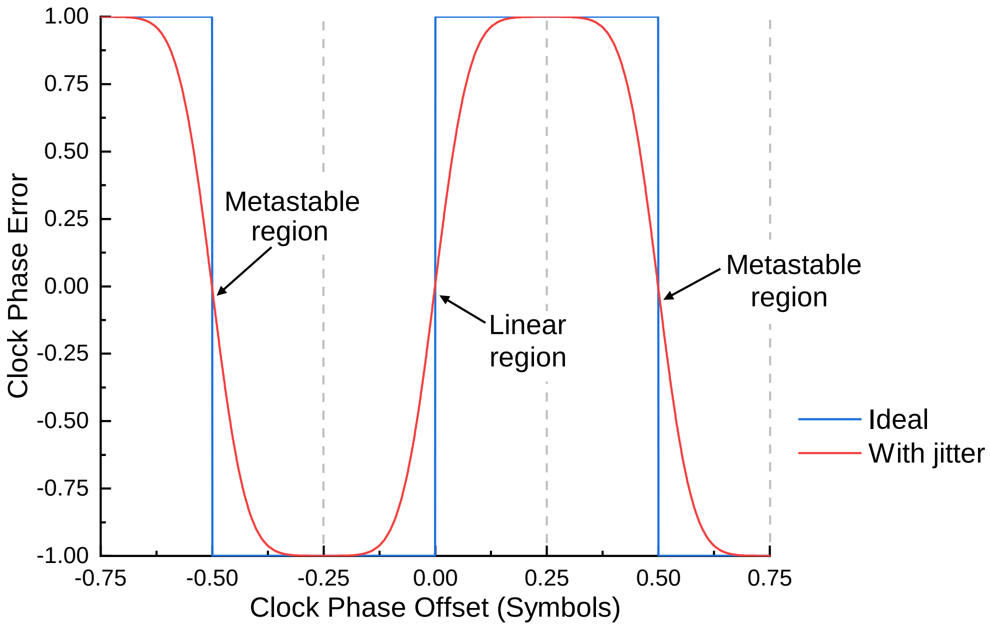

Fig. 3.4: BB-PD metastability due to jitter, a key limiter of CDR locking time in digital PI CDRs. Ideally, a BB-PD outputs a clock phase error of -1 or 1 depending on the initial clock phase offset (blue line), but jitter causes the expectation of the clock phase error to be reduced from these ideal values (red line). For initial clock phase offsets that are close to $\frac{1}{2}$ a symbol from the correct sampling position, this causes an initially slow change of the clock phase sampling position and consequently long CDR locking times.

Physically, this means that when averaged over a series of incoming data edges that have a mean clock phase that is near to $|0.5|$ symbols from the optimal sampling point (within the metastable region, e.g. at +0.49 symbols), a proportion of data edges will fall on the other side of the $|0.5|$ symbol transition (e.g. at +0.51 symbols). Data edges that arrive beyond the $|0.5|$ symbol transition have a phase error that is opposite to that generated at the mean data edge position. When averaged over many incoming data edges, this reduces the difference between the number of early and late signals, therefore reducing the movement of the CDR. Note that sampling clock jitter also causes reduction of the expectation of the clock phase error when the clock phase offset is close to the correct sampling phase in an analogous manner, which in contrast is an advantageous effect, as it reduces the movement rate of the sampling phase at the optimal sampling point. The behaviour of BB-PDs in the presence of jitter will be explored mathematically in Chapter 5.

Current research prototypes to minimise CDR locking time in PI based CDRs have focused on avoiding the metastability problem by using alternative PI-based approaches to find the correct sampling point and then handing over to a standard BB-PD [12, 69]. The current state-of-the-art PI-based CDR has an CDR locking time of 5.8 ns at 56 Gb/s (325 symbols, equivalent to 12.7 ns at 25.6 Gb/s), achieved by sweeping the phase sampling position and using a separate burst-mode phase comparator based on the exclusive OR of two successive edge samplers to find the data edge and distinguish it from the metastable position [12]. As shown in Chapter 1 Figure 1.6b, a CDR locking time of 5.8 ns (325 symbols, equivalent to 12.7 ns at 25.6 Gb/s) still results in a 35% loss of network utilisation for realistic data centre traffic versus what could be achieved with sub-nanosecond CDR locking time. Additionally, the receiver operates at a high optical power of -4 dBm [12], which is 6.5 dB higher than the minimum received power of -10.5 dBm that receivers are required to tolerate in 25 Gb/s data centre transmission standards [10].

3.2.4 Summary and Limitations of Existing Approaches

Table 3.1 contains a summary of the burst-mode CDR approaches reviewed in this section. No burst-mode CDR approach shown satisfies all requirements for data centre all-optical switching, although state-of-the-art PI based CDRs prototypes do satisfy all requirements except for CDR locking time.

Table 3.1: Comparison of burst-mode CDR approaches

Burst-Mode CDR Technique	**State-of-the-Art CDR Locking Time**	**Jitter characteristics**	**Power consumption and silicon area**	**Line rate flexibility**	**Miscellaneous**
GVCO	1-2 symbols	No jitter rejection, all jitter passed to output	Poor - multiple receivers cannot share GVCOs, each receiver CDR requires a large LC-tank GVCO	Limited to a line-rate closely matching the GVCO natural frequency	VCO natural frequency subject to PVT variation
Oversampling	Instantaneous - no need to extract clock phase from data	Large clock phase jumps between bit periods as optimum sampling point changes	Poor - multiple data samplers required and line-rate phase picking leads to high power consumption	Any line rate	Not applicable
Phase Interpolator	325 symbols (5.8 ns at 56 Gb/s) [12]	Good overall, jitter above loop bandwidth of phase tracking loop is rejected, however quantisation error of the clock phase does cause small phase jumps	Good - all digital CDR design avoids the need for each CDR to have its own dedicated VCO	Any line rate	CDR phase metastability at half a symbol initial phase offsets cause long CDR locking times.

Of these approaches, the phase interpolator based CDR is the strongest candidate CDR technology for further optimisation to achieve low CDR locking time, due to its high stability, small power consumption, and small silicon area usage [69]. However, the following two key issues currently impede the achievement of sub-nanosecond CDR locking time in this family of CDRs:

1) Asynchronous data centre networks can have frequency offsets of potentially hundreds of ppm between interconnected transceivers, due to frequency variation between different crystal oscillators located on nodes within the data centre [10]. A slow second-order loop, implemented digitally, must be used to measure and track this clock frequency offset.

2) In cases where there is no frequency offset between two interconnected transceivers, or once frequency lock is established, the initial clock phase of incoming data packets at a transceiver is entirely random. This leads to long CDR locking time due to metastability of the clock phase acquisition loop when the initial incoming clock phase of incoming data packets falls within the CDR phase detector metastable region, which occurs when data packets arrive with an approximately 0.5 symbol clock phase offset from the initial CDR sampling phase.

These two key issues could potentially be avoided through the use of frequency and phase synchronisation of transceivers interconnected through an optical switch.

3.3 Clock Synchronisation Approaches in Optical Networks

Clock time, phase and frequency synchronisation through optical fibre have been demonstrated or used in many applications, including in optical communications, in the context of clock phase synchronisation in OTDM [79, 80, 81, 82] and synchronous telecommunication standards such as SONET, SDH and Sync-E [83]. In fields outside of optical communications, example uses include in metrology [84, 85], and in time and frequency synchronisation schemes used to synchronous particle accelerator components, such as White Rabbit [16]. This section will explore some examples of synchronisation schemes that have been demonstrated in optical communications and in other fields.

3.3.1 Optical Time Division Multiplexing

In OTDM, N short (e.g. of 1 ps length) return to zero (RZ) optical pulses are multiplexed together to form a high-bandwidth pulse train, which are then demultiplexed at the receiver [86]. OTDM was heavily researched during the 1980s and 1990s for use in national telecommunications networks as a potential method of maximising per-fibre transmission rates, prior to the later introduction and use of wavelength division multiplexing (WDM) in the 2000s.

To generate the high-bandwidth pulse train, the alignment of OTDM pulses must be very precise (sub-ps). For point-to-point transmission, this can be achieved by producing ultra-fast optical pulses from the same laser, then splitting the resulting optical signal N ways with an optical coupler. Each of the N splits are separately modulated, then delayed using optical fibre delay lines of precise length, and are finally recombined into a continuous pulse train with an optical coupler [86].

Network transmission variants of OTDM were proposed in the late 1980s, where instead of generating the pulse train using one transmitter, the pulse train would be assembled at a receiver node from pulses produced by multiple transmitter nodes, with distances on metro transmission scales [80, 81, 82]. This would enable OTDM add-drop multiplexer functionality for OTDM networks, to perform the same function as the WDM add-drop multiplexers that are now used in current optical telecommunications networks. To assemble the pulse train at the receiver node, the pulse trains from a first transmitter node were clock phase aligned with pulse trains from a second transmitter node. This was achieved by measuring the clock phase offset between the two transmitters at the receiver, then generating a control signal corresponding to this clock phase offset, then transmitting this control signal back to the second transmitter, and finally using the control signal to shift its clock phase. A demonstration of this technique was performed as shown in Figure 3.5, with $\approx$50 km of fibre between each transmitter node and a fibre coupler, followed by a further $\approx$50 km between the fibre coupler and the receiver node. A BER of 10^{-9} was demonstrated with a combined receiver bit rate of 2 Gb/s [79].

Although Blank et al. [79] demonstrated that it is possible to clock phase align transmissions from multiple sources over 50 km distances, the demonstration had limitations, which were predominantly the product of the more limited technology available at the date of publication (1987) in comparison to the current state-of-the-art. These limitations are: The monitoring of the clock phase offset was achieved using a double balanced mixer [80], which are expensive, and could only be used to monitor the phase offset between a single pair of transmitters. The electronic transmitter delay was not integrated within the transmitter, and sub-nanosecond reconfiguration of the delay to switch between destinations is unlikely to have been possible as a consequence. The temperature environment the fibre was contained in was not described, though it was likely contained within a controlled laboratory environment. Data centres may experience temperature variation of up to 40 °C. The transmission rate was only 2 Gb/s, much less than current data centre data transmissions rates of at least 25 Gb/s. Significant further innovation is required to apply the concept of clock phase synchronisation to CDR in data centre optical switching.

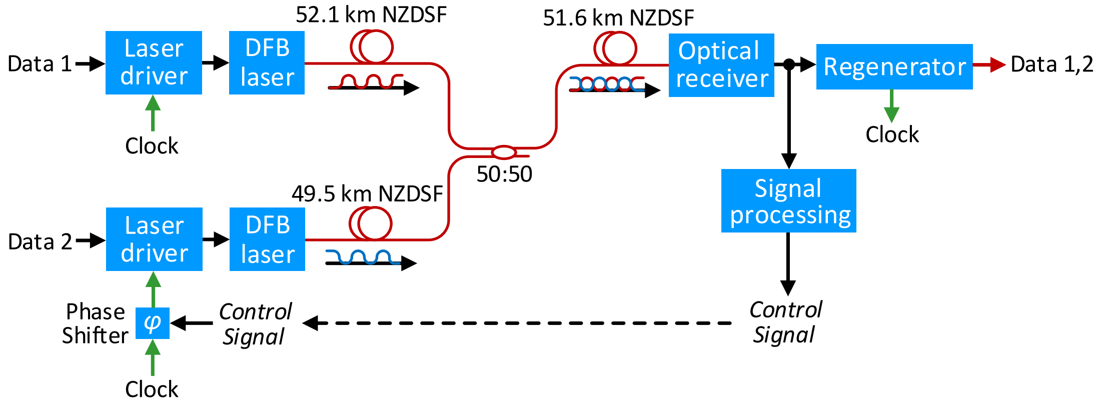

Fig. 3.5: Networked OTDM experiment. Two remote nodes generating RZ data at 1 Gb/s were time division multiplexed together to generate a combined 2 Gb/s RZ data signal at a receiving node. The clock phase of one of the two transmitters was clock phase shifted to match the second transmitter. The control signal was generated using a double balanced mixer. DFB, distributed feedback laser; NZDSF, Non-zero dispersion shifted fibre. Figure adapted from [79].

3.3.2 Synchronous Telecommunication Standards

Synchronisation within and between telecommunications networks is crucial to the reliable transmission of data across countries and continents. The Synchronous Optical Networking (SONET) [87], Synchronous Digital Hierarchy (SDH) [88] and Synchronous Ethernet (Sync-E) [83] protocols synchronise nodes in time and frequency across telecommunications networks using hierarchical clock trees, to avoid data loss from dropped packets from overflowing or underflowing receiver buffers. To ensure this when data is transmitted from one synchronised clock tree to another, which might occur when transmitting data between telecommunications networks owned by different companies, the source clock at the top of each tree must be sufficiently accurate that only one 125 μs frame slip is allowed every 72 days [83]. This is achieved using a Rubidium or Cesium atomic clock at the top of the clock tree, which maintains a time accuracy to within 1 in 10^{11}, or a Global Positioning System (GPS) informed oscillator, which keep time to within 10 μs of Coordinated Universal Time (UTC) [83]. To avoid the accumulation of jitter along the clock tree through jitter transfer, a jitter attenuator is used to remove jitter from the recovered clock at each node in the clock tree, as shown in Figure 3.6.

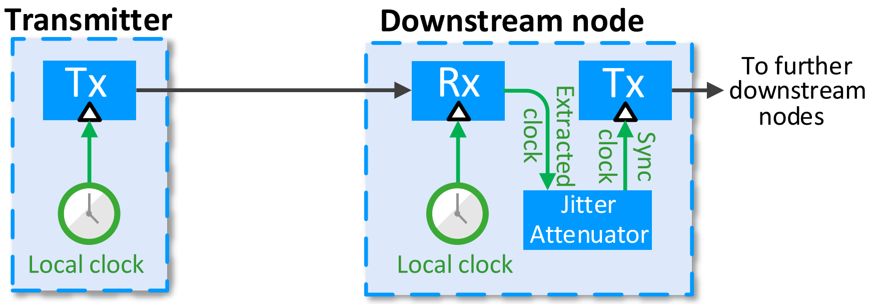

Fig. 3.6: Distribution of a clock through one hop of a SONET / SDH / Sync-E clock tree. A jitter attenuator is used to remove jitter from the recovered clock at each node in the clock tree, to minimise jitter accumulation.

Jitter attenuators have significant low frequency wander caused by environmental temperature variation, which causes the input to output delay of jitter attenuators to change. To account for this, the Sync-E standard allows up to 60 ns of low frequency wander [83]. For telecommunications networks, this is not an issue as optical links operate continuously with no burst-mode transmission. The impact of this is that receiving nodes can continuously track this wander by using large receiver buffers to compensate for the changing delay. For optically switched networks, where different nodes could experience wander in different directions due to variation in temperature throughout the data centre, the allowed low frequency wander would result in up to 120 ns of random variation within each transmitter to receiver pair. This is insufficiently accurate to perform clock phase synchronisation at 25 Gb/s, where one bit period is 40 ps in length. From a time synchronisation perspective, this would result in a 120 ns guard band being required between incoming packets to avoid collisions between packets from different transmitters. This would then dominate the optical switch reconfiguration time, reducing network utilisation to under 20% (see Figure 1.6).

3.3.3 Precision Time Protocol (PDP) and Datacentre Time Protocol (DTP)

In commercial networking, the most precise time synchronisation standard for use in asynchronous networks is the IEEE 1588 Precision Time Protocol (PTP), which is accurate to the sub-microsecond range [89]. For within data centre applications that require greater accuracy, an evolution of PTP, the Datacentre Time Protocol (DTP) protocol [90], can achieve an accuracy between synchronised nodes of under 153.6 ns, as long as there are equal to or under 6 hops between synchronised nodes [90].

To calculate the time-of-flight delay from a reference node to a downstream node, the four time points shown in Figure 3.7 are measured: t_0, the time at the reference node before transmission of the synchronisation packet; t_1, the time at the downstream node upon reception of the synchronisation packet, t_2; the time at the downstream node upon transmission of the return packet; and t_4, the time at the reference node after reception of the return packet. The time-of-flight between the nodes, Δt, is then [16]:

$$\Delta t = \frac{(t_4 - t_1) - (t_3 - t_2)}{2} \tag{3.1}$$

The calculated value of Δt is then used to adjust the time of the downstream node clock. The two critical factors limiting the accuracy of synchronisation protocols such as PTP and DTP are clock drift that occurs due to frequency mismatch between the reference and downstream node clock, and the uncertainty of measurement of Δt. The clock drift must be corrected by regularly remeasuring Δt. Factors contributing to the uncertainty of measurement of Δt include: differences in fibre length, wavelength dispersion if different wavelengths are used for the two trips and variation in latency through the serialiser-deserialiser (SERDES) transceivers. Neither PTP or DTP are sufficiently accurate for all-optically switched data centre applications, which require sub-nanosecond time synchronisation accuracy to minimise the time-synchronisation caused guard band between successive data packets.

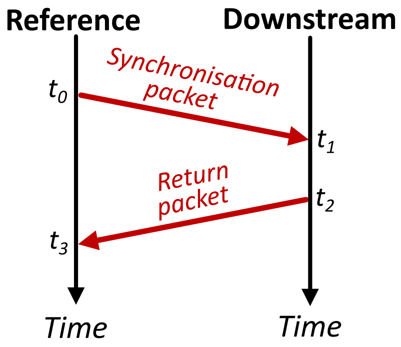

Fig. 3.7: Concept of time-of-flight measurement with PTP and DTP. Both standards operate by measuring two timestamps at the reference node, t_0 and t_3, and two timestamps at a downstream node, t_1 and t_2. Equation 3.1 is then used to calculate the time-of-flight between the reference and downstream nodes. Figure adapted from [16].

3.3.4 White Rabbit

Particle physics experiments require picosecond to nanosecond level synchronisation of measurement apparatus across distances of up to several kilometers, a similar distance scale to the data centre environment. White Rabbit combines Sync-E with PTP to synchronise measurement apparatus throughout the Large Hadron Collider (LHC) with an accuracy of nanoseconds and with picosecond precision [16], while also allowing for control signal communication through the synchronisation network. This accuracy and precision is achieved by measuring the same four time points measured in PTP, atop a network that is frequency synchronised using Sync-E, in combination with hardware improvements to minimise the differences in time-of-flight between the downstream and upstream paths. These hardware improvements include the use of: bidirectional 1 Gb/s Small Form Factor Pluggable (SFP) modules to prevent fibre length mismatches; a correction factor to account for wavelength delay differences due to fibre dispersion; measurement of SERDES transmitter latency by feeding the output of the SERDES back into a low-speed, known latency receiver in the transmitting node, and measurement of the SERDES receiver latency by feeding the data input into a low-speed, known latency receiver in the receiving node.

The accuracy of White Rabbit is sufficient for providing time synchronisation for nanosecond optical switching, and it could be used for the control plane in all-optical switches. The 1 Gb/s transmission rate is low, but fixed latency implementations of Xilinx FPGA transceivers have however since been implemented at 2.5 Gb/s using Xilinx GTP transceivers [91]. However, White Rabbit is a synchronisation network based on a tree of point-to-point links. As a consequence, though it could be used to perform time synchronisation of nodes connected to an optical switch, it could not be used to find the correct clock phase values for communicating through an optical switch without the construction of a separate clock phase calibration process.

3.4 Proposed Approach: Clock Synchronisation Assisted Clock and Data Recovery (CSA-CDR)

The two key issues for digital PI CDRs, the need to lock to any arbitrary clock frequency and the need to lock to any arbitrary clock phase, could potentially be overcome by clock synchronising transceivers that are interconnected by an optical switch, to assist and therefore simplify the CDR locking process.

The need to lock to any clock frequency could be overcome by frequency synchronising transceivers that are interconnected through an optical switch. This would eliminate the frequency offset between transceivers, eliminating the need for the slow second-order CDR frequency adaptation loop. Frequency synchronisation is appropriate in an optically-switched data centre context, since slot synchronisation is necessary in optical switches to avoid collisions between incoming data packets, which would cause data corruption [65]. Frequency synchronisation of up to 10,000 nodes connected through an optical switch is a significant challenge, but could be achieved through distribution of a central clock modulated onto a frequency comb to all nodes, as demonstrated as part of Chapter 6, or could be achieved through a distributed frequency synchronisation approach, as discussed in Ballani et al. [8].

With frequency synchronisation established, the need to lock to any arbitrary initial clock phase then still remains, incurring long CDR locking time when the initial clock phase falls within the PI metastable region. This could be overcome by phase synchronising all incoming data packets arriving at each receiver connected to an optical switch, such that the initial clock phase offset of incoming data packets never falls within the CDR metastable region, which would avoid the long CDR locking times associated with clock phase offsets that begin in this region.

This thesis names this approach clock synchronisation assisted clock and data recovery (CSA-CDR). Figure 3.8 illustrates an example optically-switched data centre architecture that uses CSA-CDR. The left of Figure 3.8 shows a typical data-centre cluster, comprising hundreds of racks (up to 64 servers per rack). An example usage case of an optical switch is then shown, which interconnects all the electronic ToR switches through an all-optical fabric, in the same fashion as proposed in optical switch architectures such as Sirius [8]. To achieve frequency synchronisation, each rack would receive a synchronised clock via a control plane, which is used to frequency synchronise the reference clock of the CDR for each transceiver. This clock could be distributed from multiple synchronised sources, so that it is tolerant to clock device failure as represented by the dotted green lines in Figure 3.8 [83, 92].

Clock phase synchronisation is established using the orange components in the right side of Figure 3.8. A clock phase cache (or store) is located within each transceiver, containing a set of values (one value per receiver) corresponding to the

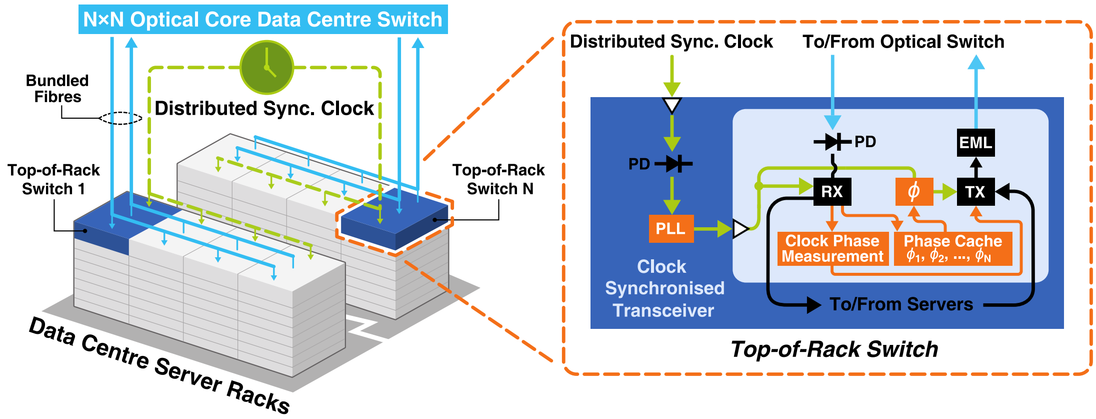

Fig. 3.8: Clock synchronised optical switch architecture with CSA-CDR. Proposed clock synchronised optically-switched data centre architecture, with CSA-CDR, which could potentially reduce CDR locking time in data centre optical switches to sub-nanosecond. Left-hand side: N data centre racks, each containing a top-of-rack switch connected to up to 64 servers, are frequency synchronised by a distributed data centre-wide synchronous clock, and are interconnected by a single $N \times N$ optical switch. Right-hand side: transceiver architecture that uses receiver clock phase measurement and transmitter clock phase shifting and storage to establish clock phase synchronisation of all transmitter to receiver paths. PLL, phase locked loop; Tx, transmitter; Rx, receiver; ϕ, clock phase interpolator; EML, externally modulated laser; MZM, Mach-Zehnder modulator; PD, photodiode. Figure adapted from [1].

phase shifts that need to be applied to the synchronised clock with a clock phase interpolator before each packet is sent from its transmitter. These clock phase values are chosen to ensure that there is only a small constrained clock phase offset when packets arrive at receivers, irrespective of packet origin. At start-up every transmitter sends a packet to all receivers connected through the optical switch in the data centre. Every receiver then measures the phase offset of each received packet and feeds back this information to the transmitters to populate their clock phase caches. These clock phase values are then used for each subsequent transmission. The resulting clock phase synchronisation of all nodes is analogous to the global clock synchronisation of transistors within ASICs.

There could then be a variety of interrelated approaches to CSA-CDR, which will be explored in the remainder of this thesis. Firstly, the clock phase values for all transmitter to receiver pairs could be calibrated only once, on initial optical switch startup (explored in Chapters 4 and 5), or the clock phase values could be regularly updated (explored in Chapter 6). Secondly, if there is only an initial clock phase calibration, the CDR in each receiver could be switched off after the initial calibration to reduce power consumption (explored in Chapter 4) or the CDR in each receiver could be used to track clock phase for incoming data packets (explored in Chapter 5). Thirdly, SMF-28 could be used for clock and data transmission (explored in Chapters 4, 5 and 6, or low thermal sensitivity HCF could be used for clock and data transmission (explored in Chapter 7).

Chapter 4

Single Calibration CSA-CDR
Part 1: Without Packet Clock Phase Tracking

4.1 Introduction

This chapter, along with Chapter 5, will explore a single calibration approach to CSA-CDR. In this approach, only a single initial calibration of clock phase values for each transmitter to receiver pair would be performed, and after this there would be no subsequent updates of the clock phase offset values. The clock phase of arriving data would then vary as data centre temperature causes fibre time-of-flight changes, causing degradation in bit error probability as the clock phase offset for each transmitter to receiver pair moves away from the optimum sampling phase.

This chapter will consider the case where the receiver CDR circuits are only be used to perform the initial clock phase calibration step. After this, the CDR circuits in each receiver would be turned off, i.e. there would be no active clock phase tracking for arriving packets. This could potentially reduce receiver power consumption as a consequence. An analytical model of receiver bit error probability, accounting for fibre time-of-flight change due to temperature, receiver noise, clock jitter and pulse shape will be established in this chapter, which will be used to evaluate the feasibility of the approach with SMF-28 fibre. This analytical model will be used as a basis for the analytical modelling that follows in Chapters 5, 6 and 7, and is presented for the first time in this thesis. Chapter 5, in contrast to this chapter, will explore the case where the CDR circuits would track the phase of incoming data packets after the initial clock phase offset calibration step.

4.2 Impact of Data Centre Environmental Conditions

If the architecture in Chapter 3 Figure 3.8 measured and applied the correct clock phase shifts under static physical environmental conditions, with no temperature change and consequently constant optical fibre time-of-flight, this would result in packets arriving at each receiver with an identical clock phase, equal to the receiver sampling clock phase, irrespective of origin transmitter. This would result in optimum bit error probability as the clock phase would never shift from the ideal receiver clock phase sampling position.

However, it is not practically or economically feasible to achieve static physical environmental conditions in the data centre. Changes in data centre temperature cause time-of-flight variation of the clock and data transmission fibres, which causes the clock phases of packets arriving at receivers to shift. If the clock phase shifts are sufficiently large, the data would be sampled away from the optimal sampling point within the incoming data signals, which would result in a poorer bit error probability. To illustrate this further, at data scales, temperature change across (for example) 2 km of optical fibre leads to fibre time-of-flight change on the order of 80 picoseconds per °C [21]. Across data-centre temperature ranges of up to 40 °C [22] and data centre distance scales of 2 km, this, in the worst case, leads to clock phase changes on the order of nanoseconds. In contrast, in ASICs, temperature change causes sub-picosecond clock delay change [93], and so no compensation for clock phase with temperature is required. If this change can be compensated for through clock phase synchronisation, synchronisation of optical switches in a manner analogous to the synchronisation of transistors within ASICs could potentially be achieved.

Additionally, noise, resulting from optical and electronic impairments, and jitter, resulting from imperfections in the quality of received embedded data clocks and receiver sampling clocks, could also cause degradation in bit error probability. Additional optical power would be required to compensate for these effects, and achieve the same bit error probability.

4.3 Analytical Modelling of Bit Error Probability Degradation from Clock Phase Shift

An analytical model of bit error probability at a receiver connected to an optical switch using single calibration CSA-CDR without packet clock phase tracking will now be derived. This analytical model will include the effect of changing data centre temperature on fibre time-of-flight, in addition to the effect of receiver electrical bandwidth, clock jitter, optical receiver noise, limited optical extinction ratio and average received optical power.

4.3.1 Sources of Signal Impairment in Intra-Data Centre NRZ-OOK Transmission

This subsection will introduce various sources that affect signal quality in NRZ-OOK optical links that are used for intra-data centre transmission. These sources will be introduced qualitatively here. Later in this chapter, they will be modelled quantitatively as they are used in the process of constructing an analytical model of an optically-switched system that uses CSA-CDR. Figure 4.1 shows an illustration of these sources.

- Optical receiver noise: noise that arises in the receiver p-i-n junction (PIN) photodiode and amplifying electronics such as a transimpedance amplifier (TIA). The dominant contributors are shot noise (which arises from quantisation of the signal photocurrent) and thermal noise (which arises from electronic noise introduced by amplifying electronics, such as a resistor or a TIA). Increased channel bandwidth increases the amount of optical noise.

- Intersymbol interference (ISI): interference of neighbouring symbols that causes signal degradation at the current data sampling point. The amount of ISI is determined by the quality of the signal pulse shape, which in turn is determined by the bandwidth of the channel. Increased channel bandwidth decreases the amount of ISI.

- Clock jitter: a combination of random and deterministic deviations of the data sampling position, which causes bit error probability to increase due to sampling at poorer quality regions of the pulse shape. Clock jitter arises from multiple deterministic and random processes within oscillators and circuits used to generate and manipulate (such as divide and multiply) the sampling clock.

- Average received optical power: the average optical power incident on the receiver photodiode. Decreased optical power received by a PIN photodiode reduces the ratio of signal photocurrent to noise photocurrent.

- Extinction ratio: the ratio of received the average optical power during a received 1 to the average received optical power during a received 0. Decreased extinction ratio decreases the differences in signal current between the 1 and 0 levels, increasing the effect of the noise photocurrent.

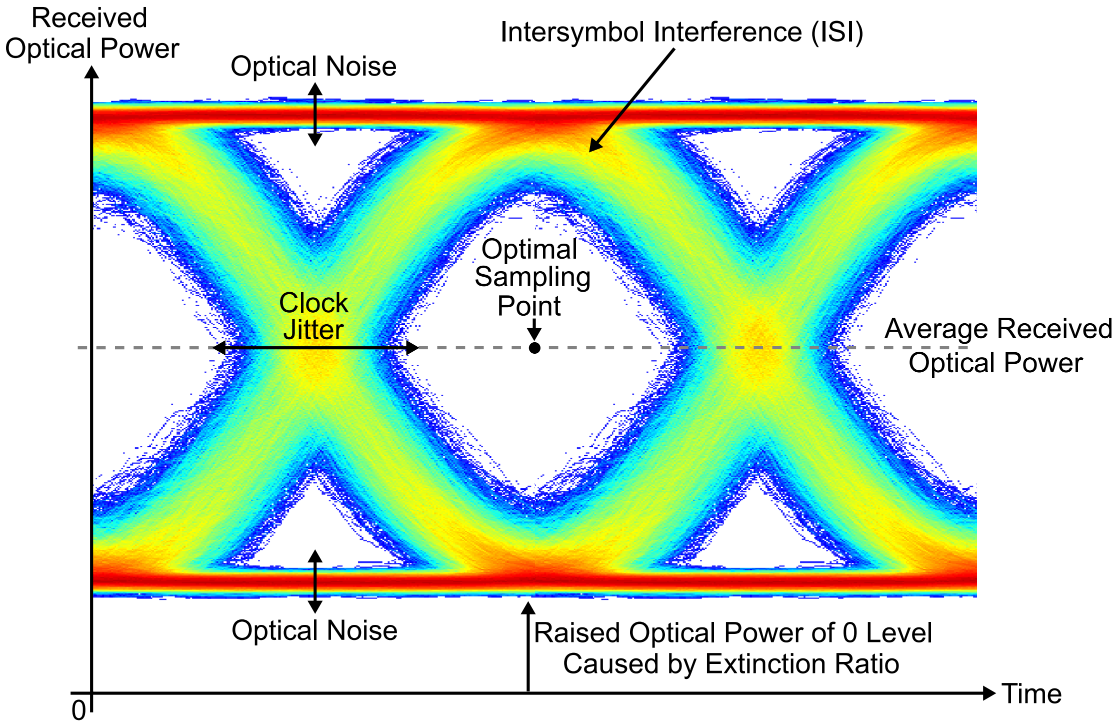
Received Optical Power
Optical Noise
Intersymbol Interference (ISI)
Optimal Sampling Point
Clock Jitter
Average Received Optical Power
Optical Noise
Raised Optical Power of 0 Level Caused by Extinction Ratio
0
Time

Fig. 4.1: Sources of signal impairment in intra-data centre NRZ-OOK transmission. The optimal sampling point is also shown. The primary source of signal impairment in this simulated NRZ-OOK signal is randomly distributed clock jitter.

4.3.2 Effect of Temperature on Fibre Time-of-Flight and Clock Phase

An increase in the temperature of a fibre causes a proportional increase in the time-of-flight of signals that propagate down that fibre, due to its thermal sensitivity. This increase results from two additive effects: the increase of fibre refractive index with increased temperature, and the physical length elongation of the fibre with temperature increase. An analysis of this effect will now be established. Firstly, the time-of-flight, t, through an optical fibre is:

$$t = n_g L/c \tag{4.1}$$

where n_g is the group refractive index of the fibre, L is the length of the fibre and c is the speed of light in vacuum. The increase in fibre time-of-flight as temperature, T, increases, $\frac{dt}{dT}$, is further given by [21]:

$$\frac{dt}{dT} = \frac{1}{c}\left(n_g \frac{dL}{dT} + L\frac{dn_g}{dT}\right) \tag{4.2}$$

The first term gives the effect of physical length elongation. The second term gives the effect of variation in group refractive index from the thermo-optic and elasto-optic effects. For the fused silica constructed SMF-28 fibres that are typically used in data centre fibres, $\frac{dL}{dT}$ = 4.1$\times$10^{-7} m/°C (for 1 m fibre) and $\frac{dn}{dT}$ = 1.1$\times$10^{-5} °C^{-1}. If n, the fibre refractive index, is used to approximate n_g, the negligible effect of glass dopants within the core is ignored and the increase in time-of-flight is calculated per fibre unit length by dividing by an additional factor of L, then the contributions of these two terms are [21]:

$$\frac{n}{cL}\frac{dL}{dT} = 2 \text{ ps/(km} \cdot^\circ \text{C)} \tag{4.3}$$

$$\frac{1}{c}\frac{dn}{dT} = 37 \text{ ps/(km} \cdot^\circ \text{C)} \tag{4.4}$$

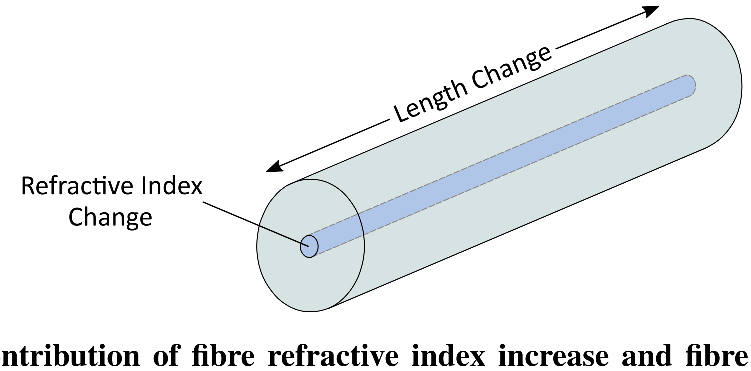

Fig. 4.2: Contribution of fibre refractive index increase and fibre expansion to increasing fibre time-of-flight with temperature. The refractive index increase is the dominant effect, contributing 95% of the overall increase in fibre time-of-flight with increased temperature.

The total contribution of these two factors is 39 ps/(km·°C), and is known as the fibre thermal coefficient of delay (TCD), τ. The analytical TCD value for SMF-28 closely matches experimental measurements of 250 μm tight buffered SMF-28 showing a TCD of 37.4 ps/(km·°C) [21] and 37.5 ps/(km·°C) [94]. Table 4.1 [94] summarises TCD values for different types of jacketed SMF-28 fibres. All types are used within data centre environments, although longer core / core distances are typically served by 250 μm or 900 μm buffered fibre bundled in ribbons within high-density cables to maximise space efficiency.

Table 4.1: Thermal coefficients of delay for different single mode fibre jacket types [94]

Fibre Type	**Thermal Coefficient of Delay (ps/(km·°C))**
250 μm tight buffered	37.5
Loose Tube	42.6
900 μm semi-tight tube	53.9
3 mm semi-tight tube	128.3

Fibre TCD (τ) has been shown to be approximately linear for 250 $\mu\mathrm{m}$ tight buffer and loose tube fibre jackets within the worst-case standardised data centre temperature range of 5 to 45 °C [22, 94]. As a consequence of this linearity, the change in time-of-flight, Δt, due to temperature change, ΔT, through these types of commonly used optical fibre may be calculated as follows:

$$\Delta t = \tau L \Delta T \tag{4.5}$$

If we define the phase, ϕ, to be the time in terms of symbols, where B is the symbol rate:

$$\phi \triangleq tB \tag{4.6}$$

then the clock phase shift, $\Delta\phi$, resulting from change in fibre time-of-flight due to temperature change, is dependent on the symbol rate of the transmitted data, B, and is [21]:

$$\Delta\phi = \tau L B \Delta T \tag{4.7}$$

To illustrate the magnitude of fibre delay changes from temperature change, consider a 2 km optical fibre, a typical maximum within-data centre building optical fibre length, undergoing temperature change of up to a worst-case data centre temperature shift of 40 °C for that length. As shown in Figure 4.3, the worst case time-of-flight change across 2 km of fibre experienced across the full 40 °C wide data centre temperature range is approximately 3 to 4 ns for 250 μm tight buffered fibre, 900 μm semi-tight tube and loose-tube buffered fibre, and approximately 10 ns for 3 mm semi-tight tube buffered fibre. These changes in time-of-flights are much larger than a single symbol period for typical current data centre transmission symbol rates of 25 GBaud. Irrespective of fibre buffer type, for a 25 GBaud NRZ-OOK signal, the worst-case time-of-flight shift is at least 75 symbols.

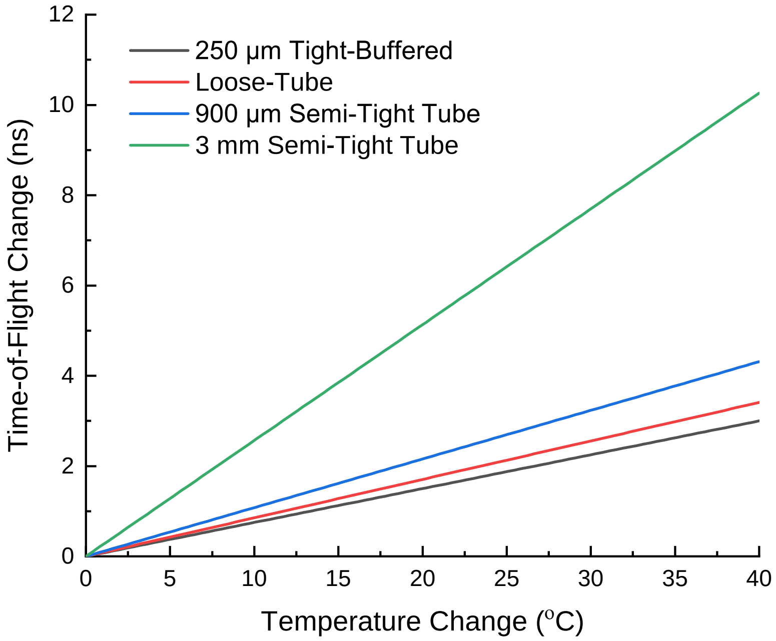

Fig. 4.3: Time-of-flight changes experienced by 2 km SMF-28 fibre with change of temperature, for a variety of typical fibre buffer types used in a data centre environment.

4.3.3 Effect of Fibre Topology on Clock Phase Offset

Consider the following transmission system, illustrated in Figure 4.4 (a single transmitter to receiver pair subset of Chapter 3 Figure 3.8): a central clock source synchronises two data centre transceivers, a transmitter and a receiver, which communicate with each other via a data centre optical switch (co-located with the clock source). This transmission system represents a single clock phase and frequency synchronised transmitter to receiver pair, which can be used to evaluate the worst-case clock phase shifts that could be experienced under worst-case changes of data centre environmental conditions.

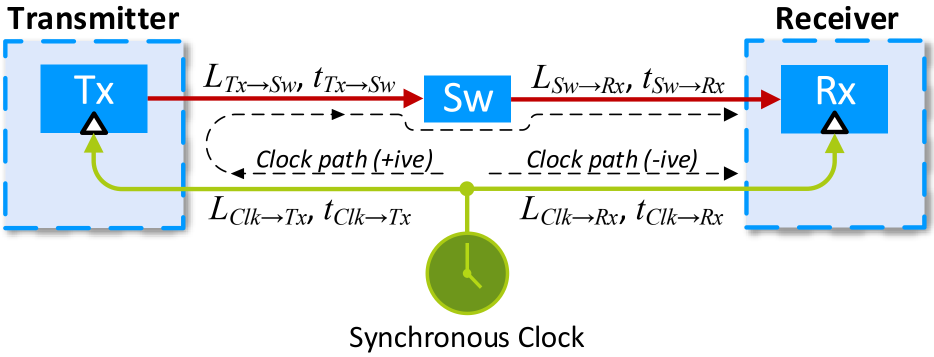
Transmitter
Receiver
Tx
Sw
Rx
$L_{Tx\to Sw}$, $t_{Tx\to Sw}$
$L_{Sw\to Rx}$, $t_{Sw\to Rx}$
Clock path (+ive)
Clock path (-ive)
$L_{Clk\to Tx}$, $t_{Clk\to Tx}$
$L_{Clk\to Rx}$, $t_{Clk\to Rx}$
Synchronous Clock

Fig. 4.4: A single pair of nodes, a transmitter and a receiver, synchronised by a central clock source and interconnected through an optical switch. Each path contributes to the overall clock phase offset experienced at the receiver. A delay increase through the clock to receiver path causes a decrease in clock phase offset (negative clock path). A delay increase through all other paths causes an increase in clock phase offset (positive clock path). Tx, transmitter; Rx, receiver; Sw, Optical Switch.

The clock source communicates a clock to the transmitter through an optical fibre of length $L_{Clk\to Tx}$ and of time-of-flight $t_{Clk\to Tx}$. The clock source also communicates a clock to the receiver through an optical fibre of length $L_{Clk\to Rx}$ and of time-of-flight $t_{Clk\to Rx}$. The transmitter communicates data to the optical switch through an optical fibre of length $L_{Tx\to Sw}$ and of time-of-flight $t_{Tx\to Sw}$. Lastly, the optical switch communicates data to the receiver through an optical fibre of length $L_{Sw\to Rx}$ and of time-of-flight $t_{Sw\to Rx}$. Assuming that the same type of optical fibre would be used for each of these lengths, each of these fibres has a TCD of τ.

Following correct calibration of the clock phase values stored within the transmitter at the beginning of operation, the clock phase offset, $\Delta\phi_{Rx-Tx}$, at the receiver for that pair, would be 0 symbols. Increases of time-of-flight through fibres $Clk\to Tx$, $Tx\to Sw$ and $Sw\to Rx$ would result in an increase in the clock phase offset at the receiver for the pair, as incoming data at the receiver would arrive later in time. Increases of time-of-flight through fibre $Clk\to Rx$ would result in a decrease in the clock phase offset at the receiver for the pair, as this would cause data to be sampled later in time.

The net change in $\Delta\phi_{Rx-Tx}$ would therefore be:

$$\Delta\phi_{Rx-Tx} = B(\Delta t_{Clk\rightarrow Tx} - \Delta t_{Clk\rightarrow Rx} + \Delta t_{Tx\rightarrow Sw} + \Delta t_{Sw\rightarrow Rx}) \tag{4.8}$$

A change in data centre temperature will cause a proportional change in these fibre time-of-flights, given by Equation 4.5. Assuming that data centre temperature increases by ΔT from the temperature at which the optical switch phase values were calibrated, the resulting change in clock phase will be:

$$\Delta\phi_{Rx-Tx} = \tau B\Delta T(L_{Clk\rightarrow Tx} - L_{Clk\rightarrow Rx} + L_{Tx\rightarrow Sw} + L_{Sw\rightarrow Rx}) \tag{4.9}$$

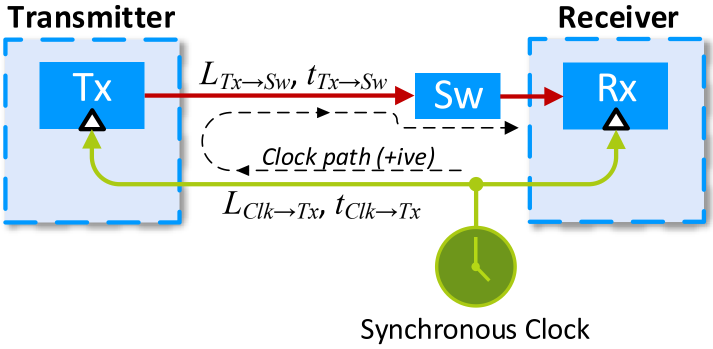
Transmitter
Receiver
Tx
Sw
Rx
$L_{Tx\rightarrow Sw}$, $t_{Tx\rightarrow Sw}$
Clock path (+ive)
$L_{Clk\rightarrow Tx}$, $t_{Clk\rightarrow Tx}$
Synchronous Clock

Fig. 4.5: Worst-case clock phase shift from co-location of the synchronous clock, the receiver and the switch. This topology eliminates the negative clock path leaving only the positive clock path. As a consequence, no partial cancellation of the clock phase offset occurs, which leads to the worst-case shift of the clock phase offset at the receiver. Tx, transmitter; Rx, receiver; Sw, Optical Switch.

The maximum possible value of $\Delta\phi_{Rx-Tx}$ occurs when there is no cancellation from $\Delta t_{Clk\rightarrow Rx}$, i.e. when $L_{Clk\rightarrow Rx}$ = 0. Practically, this would occur when the transmitter is located far from the clock source and far from the co-located optical switch, with the receiver located adjacent to the clock source and optical switch, which additionally means that $L_{Sw\rightarrow Rx}$ = 0. Under these conditions, the topology shown in Figure 4.4 reduces to the topology for worst-case clock phase shift shown in Figure 4.5. Assuming that the fibres carrying the clock and the data are both the same worst-case length, L, (causing the largest clock phase change): L = $L_{Clk\rightarrow Tx}$ = $L_{Tx\rightarrow Sw}$) and the worst-case clock phase change is given by:

$$\Delta\phi_{Rx-Tx} = 2\tau LB\Delta T \tag{4.10}$$

4.3.4 Gaussian NRZ Pulse Shape

An ideal NRZ pulse shape is square, i.e. when transmitting a 1, the transmitted signal is equal to the maximum pulse amplitude throughout the entire symbol period, and when transmitting a 0, the transmitted signal is equal to 0 throughout the entire symbol period. However, in a practical system bandwidth limitations (and other effects, such as jitter) cause degradation from this ideal pulse shape. The effect of bandwidth limitation can be modelled analytically by applying a filter to an ideal square NRZ pulse [95][†]. This subsection will derive an analytical model of a Gaussian NRZ pulse shape using this method, following the example shown in Bottacchi [95]. This will allow the error probability (expectation value of the BER) as a function of clock phase offset to be calculated. This in turn will later allow the effect of changing clock phase due to temperature change on error probability to be assessed.

An ideal square NRZ pulse, $w(\phi)$, with unit maximum pulse amplitude and unit pulse length, can be defined as [95]:

$$w(\phi) \triangleq \begin{cases} 1, & \text{if } |\phi| < \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \tag{4.11}$$

A Gaussian filter is convenient for the purpose of modelling the bandwidth limitations of a transmission system as it minimises rise and fall times while having no overshoot/undershoot [96] and it allows simple, closed-form expressions to be derived to describe the pulse shape [95] to assist in developing an analytical understanding of clock synchronisation assisted clock and data recovery. However, note that a Gaussian filter is not physically realisable as its impulse response is not causal, i.e. the impulse response does not decay completely to zero for large values of $\pm\phi$ [95]. This approach to modelling an NRZ pulse was preferred over alternate approaches, such as super-Gaussian pulse shapes, due to the clear relationship between channel bandwidth and performance.

It is convenient to define the cut-off frequency of a filter, f_c, as the frequency at which the magnitude of the response of the filter decreases to $1/\sqrt{2}$ of its maximum value (equivalent to a -3 dB reduction in power). For a low-pass Gaussian filter, the peak frequency response (peak of the Gaussian) is centred on f = 0 Hz. f_c then is equal to the -3 dB bandwidth of the low-pass filter. The amplitude-normalised frequency response of a Gaussian filter, $\hat{h}(t)$, with a cut-off frequency, f_c, and a constant phase delay of zero is then [95]:

$$\hat{H}(f) = \exp\left(- \left(\frac{f}{f_c}\right)^2 \frac{\ln 2}{2}\right) \tag{4.12}$$

[†] Note: the bandwidth of the analytically modelled filter is determined by the hardware components used in the communication link. Consequently, it is fixed for a given set of hardware components.

Figure 4.6 shows an example amplitude normalised Gaussian filter frequency response, with a cut-off frequency, f_c, and -3 dB bandwidth, of 18.6 GHz.

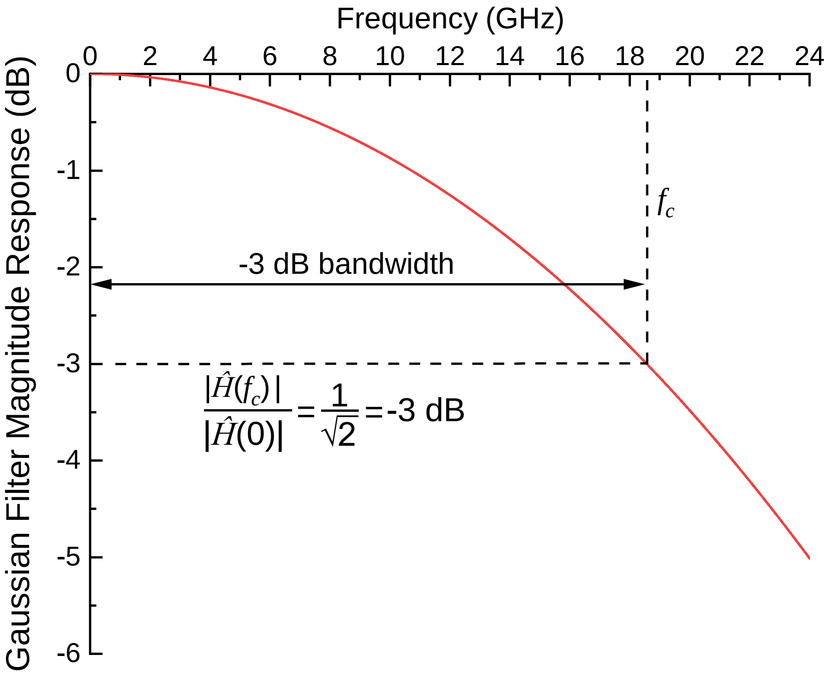

Fig. 4.6: Gaussian filter magnitude response in the frequency domain. The cut-off frequency of the Gaussian filter, f_c, is defined as the frequency at which the magnitude response of the filter decreases to $1/\sqrt{2}$ of its peak value at $f = 0$ Hz, which is equal to a -3 dB decrease in the magnitude of the response. In this example the cut-off frequency, f_c, is 18.6 GHz.

For the purpose of analytically modelling an NRZ pulse, it is convenient to perform the filtering in the time domain by convolution using the normalised filter impulse response. The normalised[†] impulse response of a Gaussian filter, $\hat{h}(t)$, can be obtained by performing the inverse Fourier transform of Equation 4.12 [95] (see Appendix A for both a full derivation and confirmation of filter normalisation):

$$\hat{h}(t) = f_c \sqrt{\frac{2\pi}{\ln 2}} \exp\left(- \left(\frac{2\pi^2 f_c^2}{\ln 2} \right) t^2 \right) \tag{4.13}$$

The standard deviation of the Gaussian filter impulse response σ_t can be calculated from the cut-off frequency f_c [95]:

$$\sigma_t = \frac{\sqrt{\ln 2}}{2\pi f_c} \tag{4.14}$$

[†] In this context, that the impulse response is normalised means that the integral of the impulse response of the filter over all time is equal to 1, which physically means that the energy of the waveform before the filter is applied is equal to the energy of the waveform after the filter has been applied.

The equivalent FWHM of the Gaussian filter impulse response, FWHM_t, can be calculated from σ_t [95]:

$$\mathrm{FWHM}_t = 2\sqrt{2\ln 2}\,\sigma_t \approx 2.35\,\sigma_t \tag{4.15}$$

The rise and fall times of the Gaussian filter impulse response are also related to σ_t. The 20–80% and 80–20% times, $\Delta t_{\text{20–80}}$ and $\Delta t_{\text{80–20}}$ respectively, are [95]:

$$\Delta t_{\text{20-80}} = \Delta t_{\text{80-20}} = \left(\sqrt{2\ln\left(\frac{1}{0.2}\right)} - \sqrt{2\ln\left(\frac{1}{0.8}\right)}\,\right)\sigma_t \approx\; 1.12\,\sigma_t \tag{4.16}$$

The standard deviation of the Gaussian filter with which the NRZ pulse shape will be derived can be defined to be in terms of symbols, k_{pls}, to simplify the mathematical derivation of the pulse shape [95]:

$$k_{\text{pls}} \triangleq \sigma_t B \tag{4.17}$$

Equations 4.14 and 4.17 can be combined to give k_{pls} in terms of f_c:

$$k_{\text{pls}} = \frac{B\sqrt{\ln 2}}{2\pi f_c} \tag{4.18}$$

The FWHM in terms of symbols, FWHM_k, can be calculated from k_{pls}:

$$\mathrm{FWHM}_k = 2\sqrt{2\ln 2}\,k_{\text{pls}} \approx 2.35\,k_{\text{pls}} \tag{4.19}$$

The rise and fall times of the Gaussian filter impulse response can also be calculated from k_{pls} and B instead of σ_t. The 20–80% and 80–20% times, $\Delta t_{\text{20-80}}$ and $\Delta t_{\text{80-20}}$ respectively, are:

$$\Delta t_{\text{20-80}} = \Delta t_{\text{80-20}} = \left(\sqrt{2\ln\left(\frac{1}{0.2}\right)} - \sqrt{2\ln\left(\frac{1}{0.8}\right)}\,\right)\frac{k_{\text{pls}}}{B} \approx\; 1.12\,\frac{k_{\text{pls}}}{B} \tag{4.20}$$

Finally, the impulse response of the filter in terms of phase, $\hat{h}(\phi)$ can be obtained by combining Equations 4.6, 4.13 and 4.18:

$$\hat{h}(\phi) = \frac{1}{k_{\text{pls}}\sqrt{2\pi}}\exp\left(-\frac{\phi^2}{2k_{\text{pls}}^2}\right) \tag{4.21}$$

An example Gaussian filter impulse response is shown in Figure 4.7. The FWHM of this impulse response is equal to 0.419 symbols at B = 25 GBaud. The cut-off frequency resulting in this impulse response is 18.6 GHz (Figure 4.6 shows the frequency response corresponding to the impulse response in Figure 4.7).

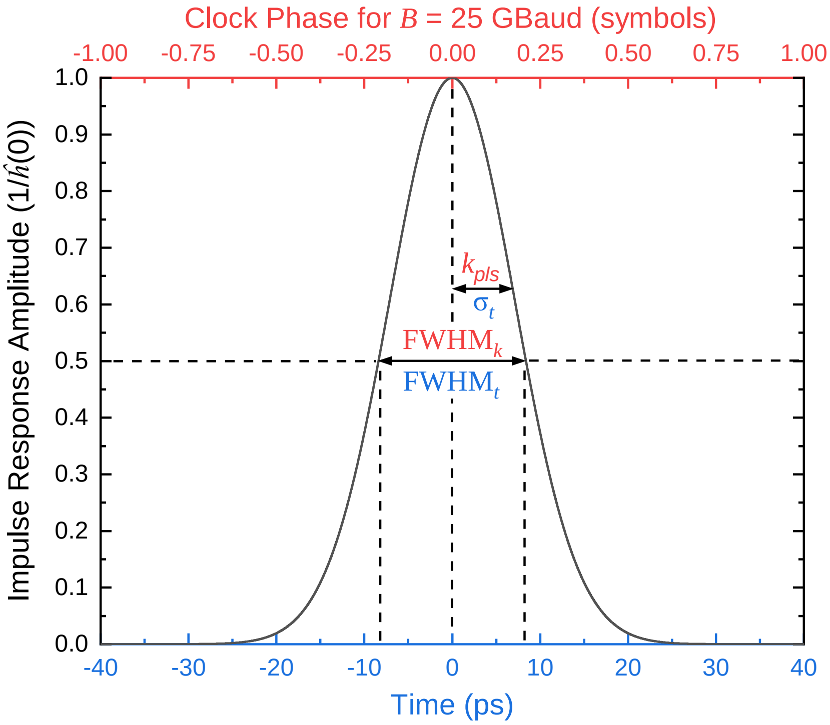

Fig. 4.7: Gaussian filter impulse response. In this example, the FWHM of the Gaussian impulse response is 0.419 symbols long for $B = 25$ GBaud (FWHM$_k$ = 0.419 symbols, FWHM$_t$ = 16.8 ps). Equivalently, in this example, k_{pls} = 0.178 symbols and σ_t = 7.12 ps. The filter cut-off frequency, f_c, that leads to this response is 18.6 GHz, which results in the Gaussian filter frequency response shown in Figure 4.6. The impulse response magnitude is adjusted such that the peak amplitude is 1 for illustrative purposes.

The pulse shape of an NRZ signal after accounting for the bandwidth limitation, modelled by a Gaussian filter response, $v(\phi)$, can be defined as the convolution of the Gaussian filter impulse response, $\hat{h}(\phi)$, with an ideal square NRZ pulse, $w(\phi)$ [95]:

$$v(\phi) \triangleq \left(\hat{h} * w\right)(\phi) \tag{4.22}$$

This convolution, using Equations 4.22, 4.21 and 4.11, can then be calculated using the following integral [95]:

$$v(\phi) = \frac{1}{k_{\text{pls}}\sqrt{2\pi}} \int_{\phi-\frac{1}{2}}^{\phi+\frac{1}{2}} \exp\left(-\frac{\alpha^2}{2k_{\text{pls}}^2}\right) d\alpha \tag{4.23}$$

The definition of the error function, $\text{erf}(z)$, is [97] (See Appendix A Figure A.1 for a plot of $\text{erf}(z)$ and the integral over a Gaussian distribution from which it is defined):

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp(-u^2) du \tag{4.24}$$

The Gaussian NRZ pulse shape can then be obtained by solving Equation 4.23 using Equation 4.24 (see Appendix A for a full derivation) [95]:

$$v(\phi) = \frac{1}{2}\left(\operatorname{erf}\left(\frac{\phi + \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right) - \operatorname{erf}\left(\frac{\phi - \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)\right) \tag{4.25}$$

The first and second $\operatorname{erf}(z)$ terms model the rising and the falling edge of the Gaussian NRZ pulse, respectively. k_{pls} is related to the rise and fall times of the Gaussian NRZ pulse edges, with larger k_{pls} resulting in longer rise and fall times.

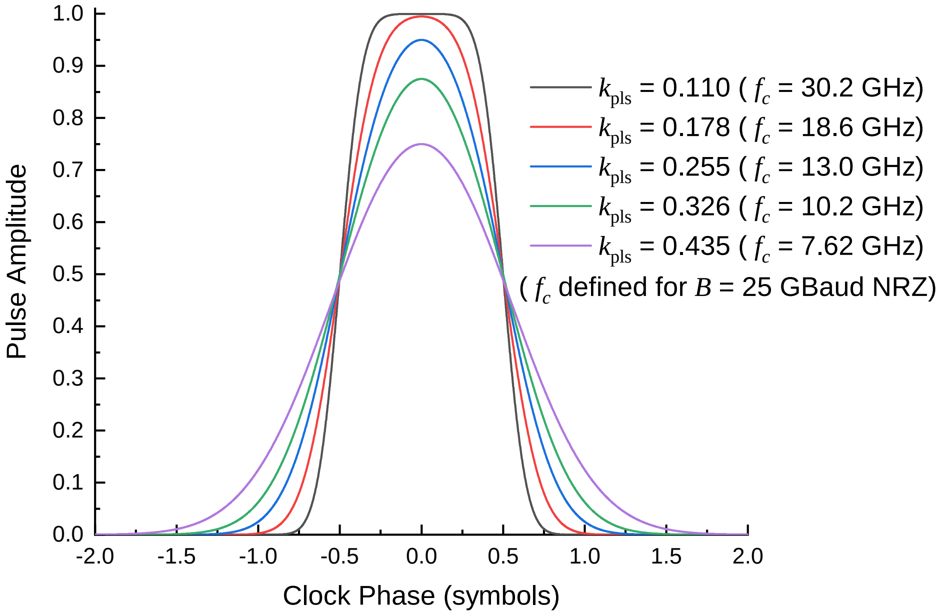

Fig. 4.8: Gaussian NRZ pulse shape. The pulse shape is calculated for a selection of values of k_{pls}, the standard deviation of the Gaussian impulse filter response. The equivalent cut-off frequencies f_c are also shown for a symbol rate of 25 GBaud.

Figure 4.8 shows the Gaussian NRZ pulse shape plotted for different values of k_{pls}. In the limit of $k_{\text{pls}} \to 0$, the Gaussian NRZ pulse shape tends towards the ideal square NRZ pulse shape given by $w(\phi)$. For $k_{\text{pls}} \gg 0$, the Gaussian NRZ pulse shape tends towards the Gaussian impulse response and the peak amplitude at $\phi = 0$, v_{peak}, decreases. The Gaussian NRZ pulse peak amplitude is directly obtained from Equation 4.25 by setting $\phi = 0$, and is given by [95]:

$$v_{\text{peak}} = \operatorname{erf}\left(\frac{1}{2\sqrt{2}k_{\text{pls}}}\right) \tag{4.26}$$

For $k_{\text{pls}} \ll 1$, $v_{\text{peak}} \approx 1$. Otherwise, increasing k_{pls} results in a decrease in v_{peak} from 1. Physically, this causes a reduction in height of the central open region of the eye due to ISI. This, as well as the amplitude of the eye opening, will now be explored in greater detail.

The Gaussian NRZ pulse shape in Equation 4.25 can be used to analytically model the eye height within the central portion of an NRZ eye as a function of phase. This is achieved analytically by first defining two pulses using Equation 4.25, a positive going pulse, $v_1(\phi)$, and a negative going pulse, $v_0(\phi)$ [95]:

$$v_1(\phi) \triangleq \frac{1}{2}\left(\operatorname{erf}\left(\frac{\phi+\frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right) - \operatorname{erf}\left(\frac{\phi-\frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)\right) \tag{4.27}$$

$$v_0(\phi) \triangleq 1 - v_1(\phi) = \frac{1}{2}\left(2 - \operatorname{erf}\left(\frac{\phi+\frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right) + \operatorname{erf}\left(\frac{\phi-\frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)\right) \tag{4.28}$$

The NRZ pulse eye opening is then defined as their difference, $v_1(\phi) - v_0(\phi)$ [95]:

$$v_1(\phi) - v_0(\phi) \triangleq \operatorname{erf}\left(\frac{\phi+\frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right) - \operatorname{erf}\left(\frac{\phi-\frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right) - 1 \tag{4.29}$$

Figure 4.9 shows $v_1(\phi) - v_0(\phi)$ for different values of k_{pls}. Note that values of $v_1(\phi) - v_0(\phi) < 0$ represent sampling at phases beyond the region where the two voltage levels cross, and so are treated as an eye height of 0.

The peak eye height, $(v_1 - v_0)_{\text{peak}}$, which occurs at ϕ = 0, can be further obtained from Equation 4.29 [95]:

$$(v_1 - v_0)_{\text{peak}} = 2\operatorname{erf}\left(\frac{1}{2\sqrt{2}k_{\text{pls}}}\right) - 1 \tag{4.30}$$

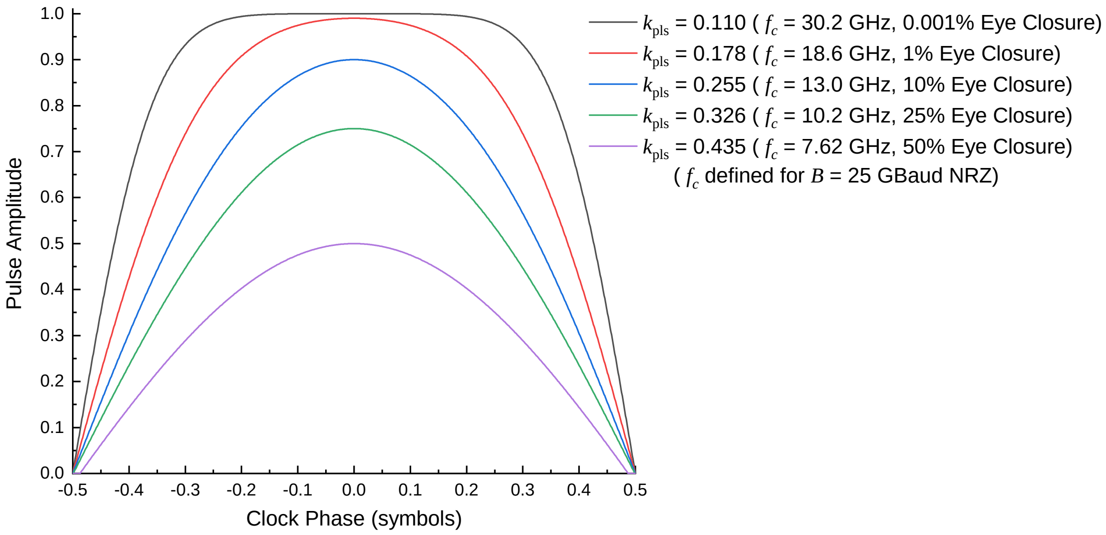

Fig. 4.9: Gaussian NRZ eye height. The NRZ eye height is calculated for a selection of values of k_{pls}, the standard deviation of the Gaussian impulse filter response. The equivalent cut-off frequencies f_c are also shown for a symbol rate of 25 GBaud, as well as the percentage reductions in eye opening height.

Figure 4.10 shows the peak eye height and the cut-off frequency as a function of k_{pls}, calculated using a symbol rate of $B = 25$ GBaud. Small k_{pls}, $k_{\text{pls}} \ll 1$, which physically represents a large cut-off frequency with respect to the symbol rate, results in a peak eye height that is approximately equal to 1. Large k_{pls}, which physically represents a small cut-off frequency with respect to the symbol rate, causes a decrease of the peak eye height from 1.

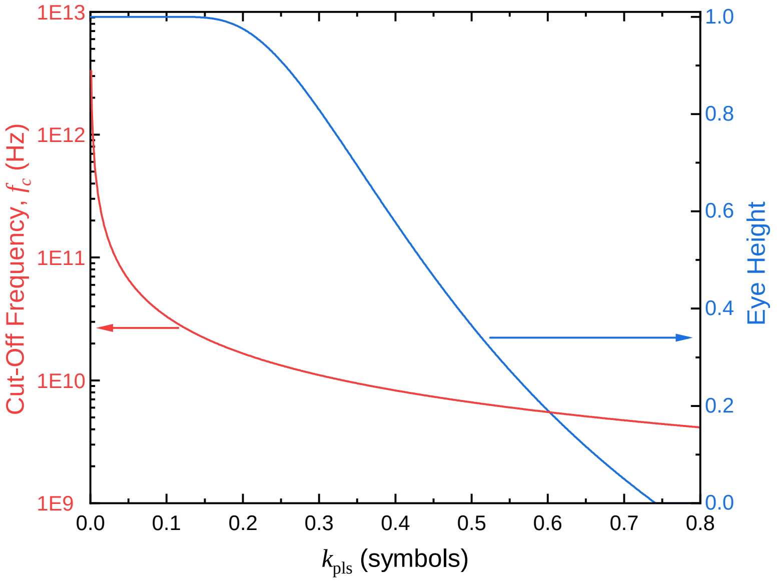

Fig. 4.10: Peak Gaussian NRZ eye height and cut-off frequency. Blue line: Gaussian NRZ peak eye height. Red line: cut-off frequency f_c. Both quantities are shown at different standard deviations of the Gaussian impulse filter response k_{pls}. A symbol rate B of 25 GBaud was used to calculate the cut-off frequency.

Sufficiently large k_{pls} also causes a decrease in the eye opening width from 1 symbol period. For $k_{\text{pls}} \ll 1$, the eye opening width is approximately equal to one symbol period. For increasing k_{pls}, as the overlap between the worst-case positive and negative pulses decreases, the eye opening width tends towards zero, at which point there is no overlap between the two worst-case pulses at all.

Large k_{pls} also results in significant ISI. Later analysis will assume that $k_{\text{pls}} \ll 1$, physically meaning that a sufficiently large cut-off frequency is used with respect to the symbol rate. This firstly ensures that the peak eye height is approximately equal to 1, secondly minimises the effect of ISI on the accuracy of the modelling, and thirdly minimises the complexity of the final equations describing the behaviour of a clock synchronisation assisted clock and data recovery system[†].

[†] A full analysis of the effect of ISI on clock synchronisation assisted CDR would add considerable complexity to the modelling that follows in this thesis and has therefore been considered out of scope. This decision was chosen as the modelling in this thesis is targeted primarily at developing an understanding of this approach to CDR, with reasonable accuracy given sufficient bandwidth to minimise ISI. A full analysis of the effect of ISI at smaller bandwidths is however an interesting potential topic for further work.

4.3.5 NRZ-OOK Photocurrent Pulse Shape Bit Error Probability

Having derived a suitable NRZ pulse shape that has a clear relationship to cut-off frequency, f_c, the next set of sub-sections will derive equations for the bit error probability of an NRZ-OOK pulse shape, modelled as the photocurrent signal and noise following optical signal reception by a PIN photodiode. These equations will allow the bit error probability at different phase offsets within an NRZ-OOK eye opening to be calculated at different incident optical powers, for typical data centre optical transmitter and receiver characteristics.

The bit error probability, p_e, when sampling an NRZ-OOK signal, is the sum of the probability of receiving a 1, $p(1)$, multiplied by the probability of erroneously sampling a 0 when a 1 should have been sampled, $P(0|1)$, and the probability of receiving a 0, $p(0)$, multiplied by the probability of erroneously sampling a 1 when a 0 should have been sampled, $P(1|0)$ [86]:

$$p_e = p(1)P(0|1) + p(0)P(1|0) \tag{4.31}$$

Equation 4.31 can be generalised to any clock phase offset within the received NRZ-OOK signal. If $P(0|1, \phi)$ is the probability of erroneously sampling a 0 when a 1 should have been sampled at a clock phase offset of ϕ, and $P(0|1, \phi)$ is the probability of erroneously sampling a 0 when a 1 should have been sampled at a clock phase offset of ϕ, then the probability of occurrence of an error when sampling the received signal, $p_e(\phi)$, at a clock phase offset of ϕ, is:

$$p_e(\phi) = p(1)P(0|1, \phi) + p(0)P(1|0, \phi) \tag{4.32}$$

In an NRZ-OOK signal, the transmitted symbols are equally likely to be a 0 or 1 due to encoding of the transmitted signal, i.e. $p(0) = p(1) = \frac{1}{2}$. The probability of an occurrence of an error when sampling the NRZ signal is then [86]:

$$p_e(\phi) = \frac{1}{2}(P(0|1, \phi) + P(1|0, \phi)) \tag{4.33}$$

The probabilities of erroneously detecting a 0 or 1, $P(0|1, \phi)$ and $P(1|0, \phi)$ respectively, then depend on the ratio of signal intensity to noise in the incoming signal. For this analysis, the signal intensity and noise will be in terms of PIN photodiode photocurrent, I, prior to TIA amplification (but including the TIAs contribution to noise). If the total noise is modelled as Gaussian distributed, which will further explored in the next section, then $P(0|1, \phi)$ and $P(1|0, \phi)$ can be calculated by performing two Gaussian integrals:

$$P(0|1,\phi) = \frac{1}{\sigma_1(\phi)\sqrt{2\pi}} \int_{-\infty}^{I_D} \exp\left(-\frac{(I(\phi)-I_1(\phi))^2}{2\sigma_1(\phi)^2}\right) dI \tag{4.34}$$

$$P(1|0,\phi) = \frac{1}{\sigma_0(\phi)\sqrt{2\pi}} \int_{I_D}^{\infty} \exp\left(-\frac{(I(\phi)-I_0(\phi))^2}{2\sigma_0(\phi)^2}\right) dI \tag{4.35}$$

where $I(\phi)$ is a the sampled photocurrent at a given phase offset in the eye, $I_D(\phi)$ is the decision threshold at that given phase offset, that delineates between $I(\phi)$ being interpreted as a 0 or a 1, $I_0(\phi)$ and $I_1(\phi)$ are the mean photocurrents of the 0 and 1 NRZ pulses respectively, and $\sigma_0(\phi)$ and $\sigma_1(\phi)$ are the variances of the 0 and 1 NRZ photocurrent pulses respectively, caused by photocurrent noise [86].

These two integrals can be solved by substitution using the definition of the complementary error function, $\mathrm{erfc}(z)$ [97]:

$$\mathrm{erfc}(z) \triangleq 1 - \mathrm{erf}(z) = 1 - \frac{2}{\sqrt{\pi}} \int_0^z \exp(-u^2) du = \frac{2}{\sqrt{\pi}} \int_z^{\infty} \exp(-u^2) du \tag{4.36}$$

$P(0|1,\phi)$ and $P(1|0,\phi)$ are then given by:

$$P(0|1,\phi) = \frac{1}{2} \mathrm{erfc}\left(\frac{I_1(\phi) - I_D(\phi)}{\sqrt{2}\,\sigma_1(\phi)} \right) \tag{4.37}$$

$$P(1|0,\phi) = \frac{1}{2} \mathrm{erfc}\left(\frac{I_D(\phi) - I_0(\phi)}{\sqrt{2}\,\sigma_0(\phi)} \right) \tag{4.38}$$

The overall probability of an error as a function of clock phase within the eye, $p_e(\phi)$, derived by substituting Equations 4.37 and 4.38 into Equation 4.33, is then:

$$p_e(\phi) = \frac{1}{4}\left(\mathrm{erfc}\left(\frac{I_1(\phi) - I_D(\phi)}{\sqrt{2}\,\sigma_1(\phi)} \right) + \mathrm{erfc}\left(\frac{I_D(\phi) - I_0(\phi)}{\sqrt{2}\,\sigma_0(\phi)} \right)\right) \tag{4.39}$$

The optimal decision level to minimise error probability, as a function of clock phase, $I_D(\phi)$, is given by [86]:

$$I_D(\phi) = \frac{\sigma_0(\phi)I_1(\phi) + \sigma_1(\phi)I_0(\phi)}{\sigma_0(\phi) + \sigma_1(\phi)} \tag{4.40}$$

4.3.6 PIN Photoreceiver NRZ-OOK Photocurrent Pulse Shape

The unit NRZ signal pulses derived earlier in this chapter will now be applied to model NRZ-OOK photocurrent pulses in PIN photoreceivers.

The signal photocurrent, I, produced by a PIN photodiode, is directly proportional to the incident optical power, P_{opt}, with a proportionality constant given by the photodiode responsivity, R [86]:

$$I = RP_{\text{opt}} \tag{4.41}$$

The maximum, average and minimum photocurrents produced by the PIN photodiode, I_{max}, I_{avg} and I_{min}, are then therefore directly proportional to the maximum, average and minimum received optical powers, P_{max}, P_{avg} and P_{min}:

$$I_{\text{max}} = RP_{\text{max}} \qquad I_{\text{avg}} = RP_{\text{avg}} \qquad I_{\text{min}} = RP_{\text{min}} \tag{4.42}$$

During transmission of a 0 in an on-off keying (OOK) signal, practical transmitters are unable to entirely prevent transmission of optical power, i.e. they have a limited extinction ratio. The relationship between the maximum optical power, P_{max}, the extinction ratio, r_e (where $r_e > 1$), and the minimum received optical power, P_{min} is:

$$P_{\text{max}} = r_e P_{\text{min}} \tag{4.43}$$

The average received optical power is the mean of the maximum and minimum received optical powers, because the probability of reception of a 0 or a 1 are both $\frac{1}{2}$:

$$P_{\text{avg}} = \frac{1}{2}(P_{\text{max}} + P_{\text{min}}) \tag{4.44}$$

The maximum and minimum received photocurrents, I_{max} and I_{min}, can then be expressed in terms of average optical power, P_{avg}, transmitter extinction ratio, r_e, and responsivity, R, using Equations 4.42, 4.43 and 4.44:

$$I_{\text{max}} = \frac{2RP_{\text{avg}}r_e}{r_e + 1} \qquad I_{\text{min}} = \frac{2RP_{\text{avg}}}{r_e + 1} \tag{4.45}$$

The mean NRZ signal positive and negative-going photocurrent pulses, $I_1(\phi)$ and $I_0(\phi)$, are then defined as the difference between the highest and lowest signal photocurrents, I_{max}, given by Equation 4.45, multiplied by the unit NRZ signal pulses, $v_1(\phi)$ and $v_0(\phi)$, defined by Equations 4.27 and 4.28. The minimum received power level arising from the limited transmitter extinction ratio is then included by adding a constant photocurrent I_{min}:

$$I_1(\phi) \triangleq (I_{\text{max}} - I_{\text{min}})v_1(\phi) + I_{\text{min}} \tag{4.46}$$

$$I_0(\phi) \triangleq (I_{\text{max}} - I_{\text{min}})v_0(\phi) + I_{\text{min}} \tag{4.47}$$

Using Equations 4.27, 4.28 and 4.45, the equations describing the NRZ signal positive and negative-going photocurrent pulses are then:

$$I_1(\phi) = RP_{\text{avg}}\left(\frac{r_e - 1}{r_e + 1}\right)\left(\text{erf}\left(\frac{\phi + \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right) - \text{erf}\left(\frac{\phi - \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)\right) + \frac{2RP_{\text{avg}}}{r_e + 1} \tag{4.48}$$

$$I_0(\phi) = RP_{\text{avg}}\left(\frac{r_e - 1}{r_e + 1}\right)\left(2 - \text{erf}\left(\frac{\phi + \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right) + \text{erf}\left(\frac{\phi - \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)\right) + \frac{2RP_{\text{avg}}}{r_e + 1} \tag{4.49}$$

The NRZ signal positive and negative-going pulses, $I_1(\phi)$ and $I_0(\phi)$, are related by:

$$I_0(\phi) = I_{\text{max}} + I_{\text{min}} - I_1(\phi) = 2RP_{\text{avg}} - I_1(\phi) \tag{4.50}$$

Worst-case average received optical power for data centre transmission links operating at 25 GBaud per lane are relatively high, which minimises error probability, thereby avoiding the need to use FEC, which is associated with an increase in power consumption and latency. Worst-case extinction ratios for data centre NRZ-OOK transmission are small. For example, the average received power in 100GBASE-LR4, standardised in IEEE 802.3, which uses 4×25 GBaud NRZ-OOK links for data centre applications, must be greater than -10.6 dBm [10]. The extinction ratio for IEEE 802.3 100GBASE-LR4 Ethernet transmitters must be at least 4 dB ($r_e >$ 2.51) [10].

Figure 4.11 gives an example pair of positive and negative NRZ-OOK photocurrent pulses following PIN photodiode reception, using an average optical power, P_{avg}, of -10.6 dBm, an extinction ratio, r_e, of 2.51 (4 dB), a Gaussian filter standard deviation, k_{pls}, of 0.189 symbols (corresponding to an NRZ symbol rate, B, of 25 GBaud and a cut-off frequency, f_c, of 17.5 GHz, matching the -3 dB bandwidth of a typical commercial TIA [98]) and a typical PIN photodiode responsivity at 1550 nm of 0.80 A/W [99].

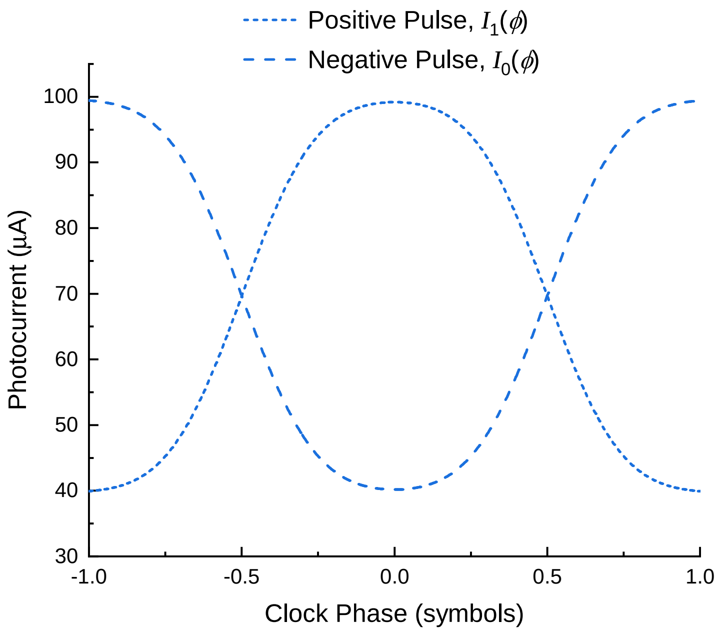

Fig. 4.11: Example pair of NRZ-OOK modulated positive and negative photocurrent pulses. The pulses are generated using a received average optical power, P_{avg}, of -10.6 dBm, following PIN reception before TIA amplification. The open region between -0.5 and 0.5 symbols is the eye opening. The extinction ratio, r_e, was modelled to be 4 dB.

4.3.7 PIN Photoreceiver NRZ-OOK Photocurrent Noise

The main contributors to noise on reception in PIN photodiode receivers will now be considered, and the total photocurrent standard deviations for the positive-going and negative-going pulses, $\sigma_1(\phi)$ and $\sigma_0(\phi)$ respectively, will be obtained.

The two contributors to photocurrent noise following OOK signal reception by a PIN photodiode are thermal noise and shot noise. Thermal noise arises from thermal agitation of charge carriers in the photodiode load resistor and amplifying electronics (typically within a TIA). Thermal noise is well described by Gaussian statistics and has a constant standard deviation, σ_T, that is independent of incident optical power. Shot noise arises from the quantisation of incident light into photons, and from the dark current, which is independent on incident optical power, and arises from the thermal generation of electron-hole charge carrier pairs in the photodiode even when no light is incident. Shot noise is most accurately modelled by Poisson statistics but is approximately Gaussian for optical communications applications, where thousands of photons are received per bit. The standard deviation of the shot noise, σ_s, increases linearly with the square root of the incident optical power, with a constant offset arising from the photodiode dark current [86].

Since thermal noise and shot noise are independent of each other, and are each modelled by random Gaussian processes, the overall variance of the receiver noise, σ^2, is the sum of the thermal noise and shot noise variances, i.e. $\sigma^2 = \sigma_s^2 + \sigma_T^2$. The overall photocurrent noise variance following PIN photodiode OOK signal reception, σ^2, can then therefore be modelled using Gaussian statistics [86].

To explore the physical origin of the variance of the thermal noise and the shot noise in a PIN photodiode in more detail, the variance of the thermal noise σ_T^2 and the variance of the shot noise, σ_s^2, can be modelled by the following equations [86]:

$$\sigma_T^2 = \frac{4k_B T F_n f_c}{R_L} = \sigma_{ni}^2 \tag{4.51}$$

$$\sigma_s^2 = 2q(I_p + I_d)f_c = 2q(RP_{\text{opt}} + I_d)f_c \tag{4.52}$$

where k_B is Boltzmann's constant (1.380649×10^{-23} m^2 kg s^{-2} K^{-1} [100]), T is the absolute temperature, F_N is the amplifier noise figure, representing the factor by which thermal noise is enhanced by various resistors used in amplifiers following the PIN, f_c is the cut-off frequency of the PIN, R_L is the resistance of the load resistor, q is the charge of one electron (1.602176634×10^{-19} C [101]), I_p is the photodiode photocurrent and I_d is the photodiode dark current. The photodiode photocurrent, I_p, can also be given in terms of received optical power, P_{opt}, using Equation 4.41, $I_p = RP_{\text{opt}}$.

The total variance of the noise in a PIN photoreceiver is then:

$$\sigma^2 = \sigma_T^2 + \sigma_s^2 = \left(2q(I_p + I_d) + \frac{4k_B T F_n}{R_L}\right) f_c \tag{4.53}$$

The total receiver photocurrent noise variances can then be split into two constant terms, σ_T^2 and σ_d^2, arising from the thermal noise and the dark current contributions to shot noise respectively, and a variable term depending on incident optical power, σ_{opt}^2, arising only from the photocurrent contribution to shot noise, I_p, which arises from the incident optical power, P_{opt}:

$$\sigma_T^2 = \frac{4k_B T F_n f_c}{R_L} = \sigma_{ni}^2 \tag{4.54}$$

$$\sigma_d^2 = 2qI_d f_c \tag{4.55}$$

$$\sigma_{\text{opt}}^2 = 2qI_p f_c = 2qRP_{\text{opt}} f_c \tag{4.56}$$

To put these photocurrent noise contributions into context, consider an example typical photoreceiver with PIN photodiode and TIA, designed for 25 GBaud OOK signal reception, with the characteristics shown in Table 4.2:

Table 4.2: Example characteristics of a typical commercial PIN photodiode receiver intended for 25 GBaud NRZ-OOK reception. The PIN photodiode receiver consists of a TIA [98] and PIN photodiode [99]. (Note: the TIA noise figure was calculated indirectly based on the TIA input referred noise, the TIA transimpedance, and room temperature of 298 K. The cut-off frequency was assumed to be limited by the TIA, as the photodiode bandwidth was 22 GHz [99].)

Photodiode / TIA property	**Value**
Photodiode responsivity, R	0.80 A/W [99]
Photodiode dark current, I_d	2 nA [99]
TIA transimpendance, R_L	7.5 kΩ [98]
TIA cut-off frequency, f_c	17.5 GHz [98]
TIA input referred noise, σ_{ni}	2 μA [98]
TIA noise figure, F_N	20.2 dB

These characteristics of a typical PIN photodiode receiver with TIA are then used to plot Figure 4.12. This figure shows the standard deviation of the total noise, σ, and its contributors from the constant terms, the thermal noise, σ_T, and the dark current shot noise, σ_d, and the term that varies with optical power, the photocurrent shot noise σ_{opt}, plotted against optical power for 25 Gb/s data centre NRZ-OOK transmission using the characteristics of a typical PIN photodiode receiver with TIA given in Table 4.2. At small optical powers, e.g. ≈ -20 dBm, the photoreceiver operates in the thermal noise limit, where the shot noise is much smaller than the thermal noise. At large optical powers, e.g. ≈ 10 dBm, the photoreceiver operates in the shot noise limit, where the thermal noise is much smaller than the shot noise. Typical standardised data centre received optical powers fall into neither limit.

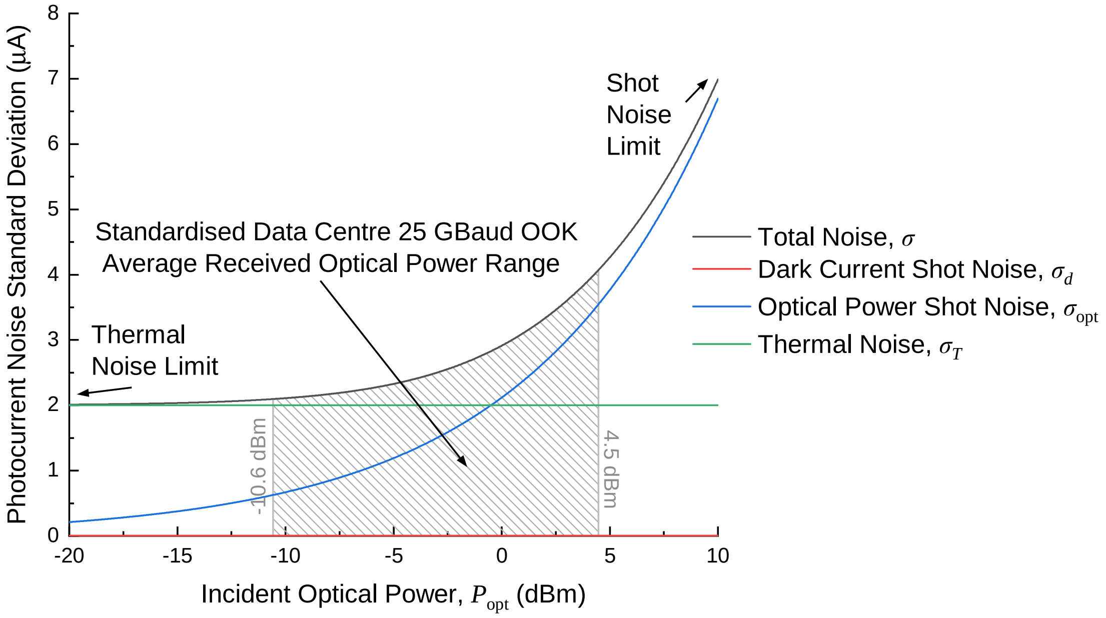

Fig. 4.12: Total PIN photocurrent noise, as well as its contributors. These quantities are plotted using the example characteristics given in Table 4.2, of a typical commercial PIN photodiode receiver intended for 25 GBaud NRZ-OOK reception. A typical operational range for data centre received optical power at 25 GBaud NRZ-OOK reception is shown [10].

The effect of the standardised data centre received optical powers falling in between the shot noise and the thermal noise limits is that, firstly, the standard deviation of the total noise cannot be approximated as constant, due to the contribution of shot noise, and secondly, the constant contribution of the thermal noise cannot be ignored. The contribution of dark current to the total noise can be neglected, as its contribution to the total noise is negligible (in Figure 4.12, its constant standard deviation is only 3.34 nA vs the constant standard deviation of the thermal noise of 2 μA). If the TIA input referred noise, σ_{ni}, is used to maximise equation simplicity, the photocurrent noise variances, $\sigma_1^2(\phi)$ and $\sigma_0^2(\phi)$, for the positive and negative-going photocurrent pulses respectively, are then:

$$\sigma_1^2(\phi) = 2qI_1(\phi)f_c + \sigma_{ni}^2 \tag{4.57}$$

$$\sigma_0^2(\phi) = 2qI_0(\phi)f_c + \sigma_{ni}^2 \tag{4.58}$$

The signal photocurrent, I_p, has been defined as a function of clock phase for the positive and negative-going pulses in the previous section in Equations 4.48 and 4.49. By substituting in for $I_1(\phi)$ and $I_0(\phi)$, the variances of the noise of the positive and negative-going pulses are then:

$$\sigma_1^2(\phi) = 2qRP_{\text{avg}}\left(\frac{r_e-1}{r_e+1}\right)\left(\text{erf}\left(\frac{\phi+\frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right) - \text{erf}\left(\frac{\phi-\frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)\right)f_c + \frac{4qRP_{\text{avg}}f_c}{r_e+1} + \sigma_{ni}^2 \tag{4.59}$$

$$\sigma_0^2(\phi) = 2qRP_{\text{avg}}\left(\frac{r_e-1}{r_e+1}\right)\left(2 - \text{erf}\left(\frac{\phi+\frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right) + \text{erf}\left(\frac{\phi-\frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)\right)f_c + \frac{4qRP_{\text{avg}}f_c}{r_e+1} + \sigma_{ni}^2 \tag{4.60}$$

where r_e is the transmitter extinction ratio and k_{pls} is the standard deviation in symbols of the Gaussian filter used to obtain the NRZ-OOK photocurrent pulse shape. k_{pls} is related to the TIA cut-off frequency, f_c, by the symbol rate, B:

$$k_{\text{pls}} = \frac{B\sqrt{\ln 2}}{2\pi f_c} \tag{4.61}$$

Further simplification of Equations 4.59 and 4.60 is not possible without sacrificing accuracy at larger optical powers within the -10.6 to 4.5 dBm typical operational range for 25 Gb/s data centre PIN optical receivers [10]. Although the bit error probability is small at large optical powers due to a high associated signal to noise ratio, inaccuracy of the pulse shape bit error probability calculation could lead to inaccuracy of the temperature range tolerance calculation that follows in a later section of this chapter. The small worst-case extinction ratio of data centre optical transmitters also does not permit further simplification of Equations 4.59 and 4.60.

Figure 4.13 shows the noise standard deviation for the example pair of NRZ-OOK modulated positive and negative photocurrent pulses shown in Figure 4.11, calculated using Equations 4.59 and 4.60, and the example typical PIN photodiode characteristics given in Table 4.2. The noise standard deviation was calculated at a received optical power of -10.6 dBm, a typical minimum optical power for 25 GBaud NRZ-OOK data centre transmission. Although the proportion of the shot noise contribution to optical noise standard deviation is relatively small in Figure 4.13, the proportion of shot noise at higher optical powers falling within the data centre received optical power range is much greater, as shown in Figure 4.12.

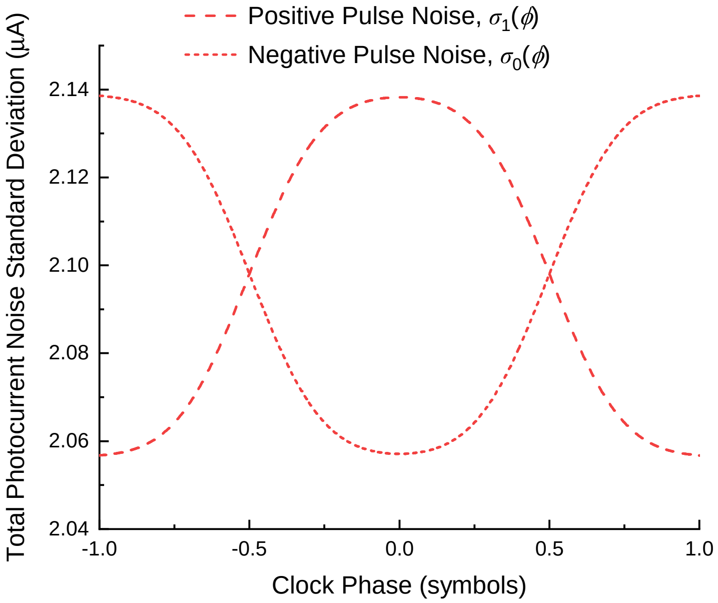

Fig. 4.13: Standard deviation of the total noise for the example pair of NRZ-OOK modulated positive and negative photocurrent pulses shown in Figure 4.11. The total noise is generated following PIN reception, prior to TIA transimpedance amplification but including the TIA input referred noise. The example PIN photodiode and TIA characteristics from Table 4.2 are used to calculate the noise.

4.3.8 NRZ-OOK Photocurrent Pulse Shape Bit Error Probability

The equations describing NRZ bit error probability (Equation 4.33), the photocurrent after PIN photodiode reception of an NRZ-OOK signal (Equations 4.48 and 4.49) and the photocurrent noise (Equations 4.57 and 4.58) can now be combined to give NRZ-OOK bit error probability, p_e, as a function of clock phase, ϕ. Combining these equations together, the bit error probability within an received NRZ-OOK signal modelled by a Gaussian NRZ pulse is then:

$$p_e(\phi) = \frac{1}{4}\left(\operatorname{erfc}\left(\frac{I_1(\phi) - I_D(\phi)}{\sqrt{4qI_1(\phi)f_c + 2\sigma_{ni}^2}}\right) + \operatorname{erfc}\left(\frac{I_D(\phi) - I_0(\phi)}{\sqrt{4qI_0(\phi)f_c + 2\sigma_{ni}^2}}\right)\right) \tag{4.62}$$

where ϕ is the clock phase offset, $I_D(\phi)$ is the photocurrent decision threshold, q is the electronic charge of one electron, f_c is the cut-off frequency of the TIA, σ_{ni} is the input referred noise of the TIA, and $I_1(\phi)$ and $I_0(\phi)$ are the positive and negative-going NRZ-OOK photocurrent pulses given in Equations 4.48 and 4.49, and are:

$$I_1(\phi) = RP_{\text{avg}}\left(\frac{r_e - 1}{r_e + 1}\right)\left(\operatorname{erf}\left(\frac{\phi + \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right) - \operatorname{erf}\left(\frac{\phi - \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)\right) + \frac{2RP_{\text{avg}}}{r_e + 1} \tag{4.63}$$

$$I_0(\phi) = RP_{\text{avg}}\left(\frac{r_e - 1}{r_e + 1}\right)\left(2 - \operatorname{erf}\left(\frac{\phi + \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right) + \operatorname{erf}\left(\frac{\phi - \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)\right) + \frac{2RP_{\text{avg}}}{r_e + 1} \tag{4.64}$$

where r_e is the transmitter extinction ratio, k_{pls} is the standard deviation in symbols of the Gaussian filter used to obtain the NRZ-OOK photocurrent pulse shape, R is the PIN photodiode responsivity and P_{avg} is the average received power. Finally, k_{pls} is related to the cut-off frequency of the TIA, f_c, by the symbol rate, B:

$$k_{\text{pls}} = \frac{B\sqrt{\ln 2}}{2\pi f_c} \tag{4.65}$$

To illustrate these equations, Figure 4.14 shows a series of receiver eye diagrams generated by evaluating Equation 4.62 against photocurrent decision threshold, $I_D(\phi)$, between the minimum and maximum pulse photocurrents at a set of different incident average optical powers, P_{avg}, at clock phase offsets, ϕ, between -0.5 and 0.5 symbols. An extinction ratio, r_e, of 4 dB, and the PIN photodiode and TIA characteristics in Table 4.2 are used, along with $k_{\text{pls}} = 0.189$ symbols (corresponding to a cut-off frequency, f_c, of 17.5 GHz at 25 GBaud, matching the TIA bandwidth). Smaller average received optical powers have more closed eye diagrams with greater error probability throughout the eye. Note that Equation 4.62 and the eye diagrams in Figure 4.14 do not include the effect of clock jitter, the effect of which will be accounted for in a later section.

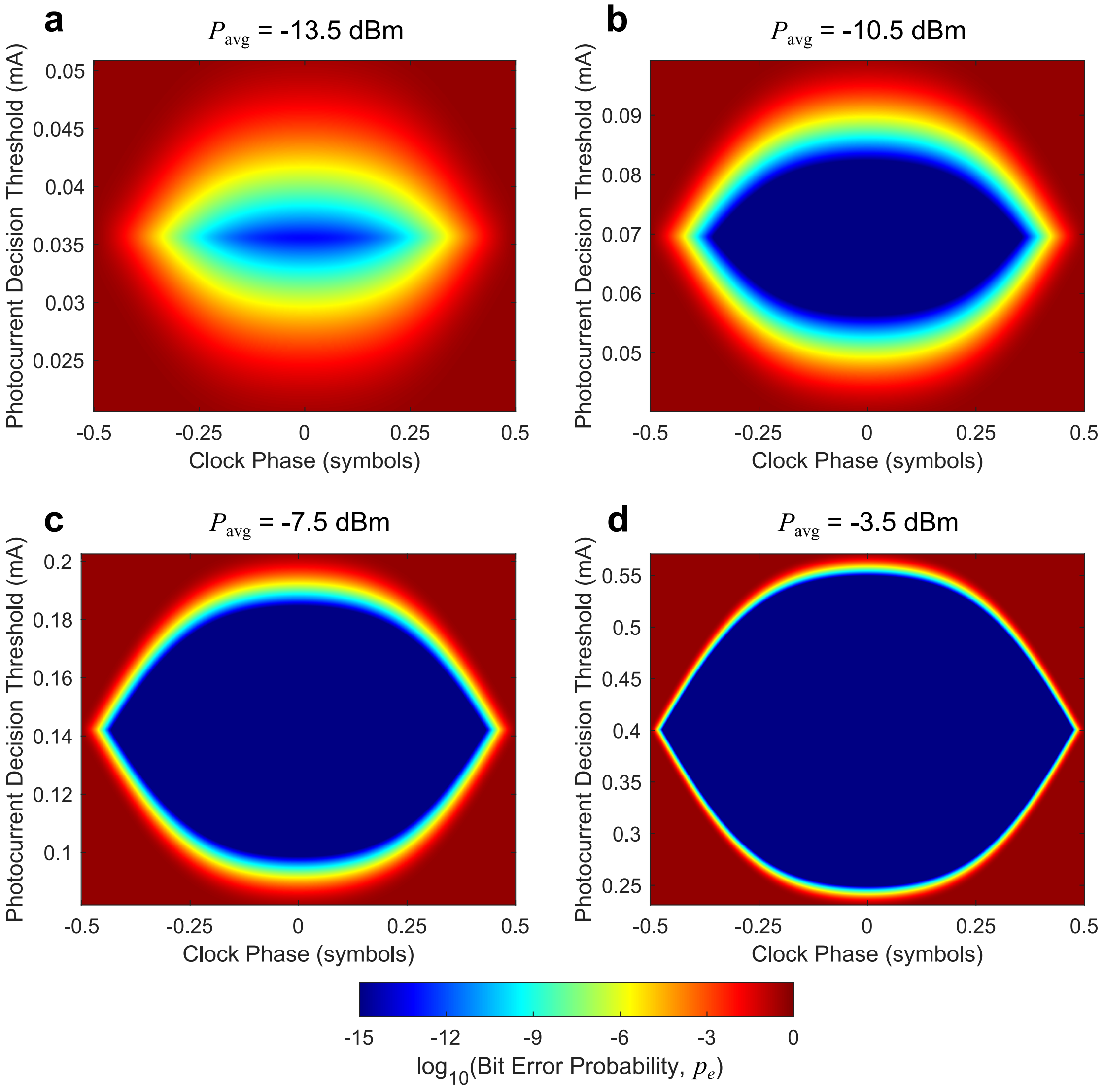

Fig. 4.14: Analytically modelled receiver eye diagrams. The error probability, p_e, is calculated using Equation 4.62, as a function of clock phase, ϕ, at a series of different average received optical powers. **a**, P_{avg} = -13.5 dBm, near the sensitivity of the PIN photoreceiver; **b**, P_{avg} = -10.5 dBm, a typical minimum received optical power for data centre 25 GBaud NRZ-OOK signals [10]; **c**, P_{avg} = -7.5 dBm and **d**, P_{avg} = -3.5 dBm, two larger average optical powers falling within a typical permitted average received optical power range of -10.6 dBm to 4.5 dBm for data centre 25 GBaud NRZ-OOK signals [10]. Error probabilities smaller than 10^{-15} are displayed as a probability of 10^{-15}.

Many receivers, such as those used in FPGAs, may not contain the necessary circuitry to account for the lower optimum photocurrent decision threshold at greater optical powers, which occurs due to a greater contribution of shot noise to the overall total noise. As a further observation, the optimum decision threshold can vary significantly as a function of phase, ϕ, in situations where shot noise is dominant in receiver noise[†]. Given that a receiver will not initially know what phase offset data arrives with, a receiver would be unable to know which decision point would be optimal to sample to data.

To account for these observations, the photocurrent decision threshold, I_D, will be defined to be equal to halfway between the minimum and maximum received photocurrents, irrespective of clock phase offset, ϕ. The probabilities of transmitting a 1 or 0 through the link are both $\frac{1}{2}$, so the photocurrent halfway between the minimum and maximum photocurrents is equal to the mean receiver photocurrent, which is in turn proportional to the average received optical power, P_{avg}. Using Equation 4.41, the photocurrent decision threshold, I_D, can then therefore be defined in the context of data centre NRZ-OOK reception as:

$$I_D \triangleq RP_{\text{avg}} \tag{4.66}$$

Once this is accounted for, the definition for I_D in Equation 4.66 can be substituted into Equation 4.62, which gives the bit error probability within an received NRZ-OOK signal modelled by a Gaussian NRZ pulse. The error probability against clock phase, $p_e(\phi)$, is then:

$$p_e(\phi) = \frac{1}{4}\left(\operatorname{erfc}\left(\frac{I_1(\phi) - RP_{\text{avg}}}{\sqrt{4qI_1(\phi)f_c + 2\sigma_{ni}^2}} \right) + \operatorname{erfc}\left(\frac{RP_{\text{avg}} - I_0(\phi)}{\sqrt{4qI_0(\phi)f_c + 2\sigma_{ni}^2}} \right) \right) \tag{4.67}$$

where q is the electronic charge of one electron, f_c is the cut-off frequency of the TIA, σ_{ni} is the TIA input referred noise, and $I_1(\phi)$ and $I_0(\phi)$ are the positive and negative-going NRZ-OOK photocurrent pulses given in Equations 4.48 and 4.49, and are:

$$I_1(\phi) = RP_{\text{avg}}\left(\frac{r_e - 1}{r_e + 1}\right)\left(\operatorname{erf}\left(\frac{\phi + \frac{1}{2}}{\sqrt{2}k_{\text{pls}}} \right) - \operatorname{erf}\left(\frac{\phi - \frac{1}{2}}{\sqrt{2}k_{\text{pls}}} \right) \right) + \frac{2RP_{\text{avg}}}{r_e + 1} \tag{4.68}$$

$$I_0(\phi) = RP_{\text{avg}}\left(\frac{r_e - 1}{r_e + 1}\right)\left(2 - \operatorname{erf}\left(\frac{\phi + \frac{1}{2}}{\sqrt{2}k_{\text{pls}}} \right) + \operatorname{erf}\left(\frac{\phi - \frac{1}{2}}{\sqrt{2}k_{\text{pls}}} \right) \right) + \frac{2RP_{\text{avg}}}{r_e + 1} \tag{4.69}$$

[†] To understand why this effect occurs, consider a heavily shot noise impacted signal. At $\phi = 0$, the total noise on the 1 level will be much greater than the noise on the 0 level, as the difference in mean photocurrent is maximised at this clock phase, resulting in an optimal decision point that is lower than $\frac{1}{2}(I_1 + I_0)$ using Equation 4.40. At $\phi = \pm\frac{1}{2}$, the mean photocurrents for the 0 and 1 levels are equal, so there is therefore no difference in the photocurrent noise magnitudes, which results in $I_D = \frac{1}{2}(I_1 + I_0)$ using Equation 4.40.

where r_e is the transmitter extinction ratio, k_{pls} is the standard deviation in symbols of the Gaussian filter used to obtain the NRZ-OOK photocurrent pulse shape, R is the PIN photodiode responsivity and P_{avg} is the average received power.

Assuming a constant decision point, I_D, of RP_{avg}, as used in Equation 4.67, the error probability as a function of clock phase, $p_e(\phi)$, can then be evaluated for a series of different average received optical powers, P_{avg}, using Equation 4.67. Figure 4.15 then shows the error probability as a function of phase offset using a decision point of RP_{avg}, for the four eye diagrams in Figure 4.14.

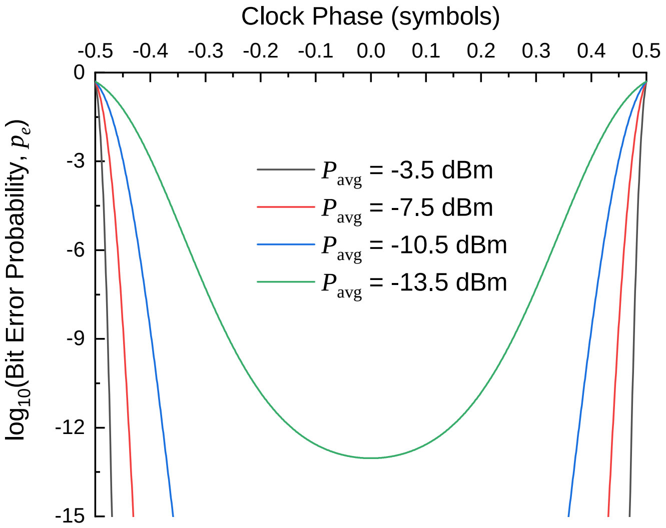

Fig. 4.15: OOK error probability, p_e, as a function of clock phase offset. The error probability was calculated assuming a Gaussian impulse response with $k_{\mathrm{pls}} = 0.189$ (corresponding to a cut-off frequency f_c of 17.5 GHz for a 25 GBaud NRZ signal), calculated using the four optical powers used to generate the eye diagrams in Figure 4.14.

4.3.9 Effect of Jitter on Clock Phase Shifted NRZ-OOK Error Probability

In practical receivers, in addition to optical noise, the position of the sampling clock with respect to the incoming data is also subject to jitter (which can be considered as noise in the time domain). With increased jitter of the sampling clock within the NRZ-OOK eye, more samples are taken in regions of the NRZ-OOK eye with a poorer signal to noise ratio. This leads to an increase in error probability. When there is no clock phase offset between the optimal sampling phase in the incoming data and the sampling clock, the error probability after accounting for jitter, $p_{e(\text{post}-\text{jit})}$ can be calculated by integrating the jitter probability distribution function (PDF), PDF_{jit}, with the error probability resulting from the eye shape, p_e [95]:

$$p_{e(\text{post}-\text{jit})} = \int_{-\infty}^{\infty} \text{PDF}_{\text{jit}}(\phi) p_{e(\text{pre}-\text{jit})}(\phi)\, d\phi \tag{4.70}$$

The error probability after accounting for jitter as a function of clock phase offset can be calculated by extending the integral in Equation 4.70 to a cross-correlation between PDF_{jit} and p_e:

$$p_{e(\text{post}-\text{jit})}(\phi) = (\text{PDF}_{\text{jit}} \star p_{e(\text{pre}-\text{jit})})(\phi) \tag{4.71}$$

This cross-correlation may also be written as:

$$p_{e(\text{post}-\text{jit})}(\phi) = \int_{-\infty}^{\infty} \text{PDF}_{\text{jit}}(\alpha) p_e(\phi + \alpha)\, d\alpha \tag{4.72}$$

In practical systems, clock jitter contains contributions from both random jitter (which arises from effects such as thermal noise) and from deterministic jitter (which arises from effects such as duty cycle distortion). To preserve simplicity in this analytical modelling, clock jitter will be treated as entirely random, which can be modelled with a Gaussian distribution with a standard deviation of σ_{jit}.

It is convenient to define the Gaussian jitter magnitude in terms of symbols, with a standard deviation of k_{jit}:

$$k_{\text{jit}} \triangleq \sigma_{\text{jit}} B \tag{4.73}$$

The PDF of the jitter, PDF_{jit}, modelled by a Gaussian distribution with a mean of 0 symbols and standard deviation k_{jit}, is then:

$$\text{PDF}_{\text{jit}}(\alpha) = \frac{1}{k_{\text{jit}}\sqrt{2\pi}} \exp\left(- \frac{\alpha^2}{2k_{\text{jit}}^2}\right) \tag{4.74}$$

Figure 4.16 shows a set of example Gaussian distributions modelling random jitter, with a series of different jitter standard deviations defined terms of both symbols at 25 GBaud and time. Greater jitter standard deviations cause a greater magnitude of random variation of the sampling clock.

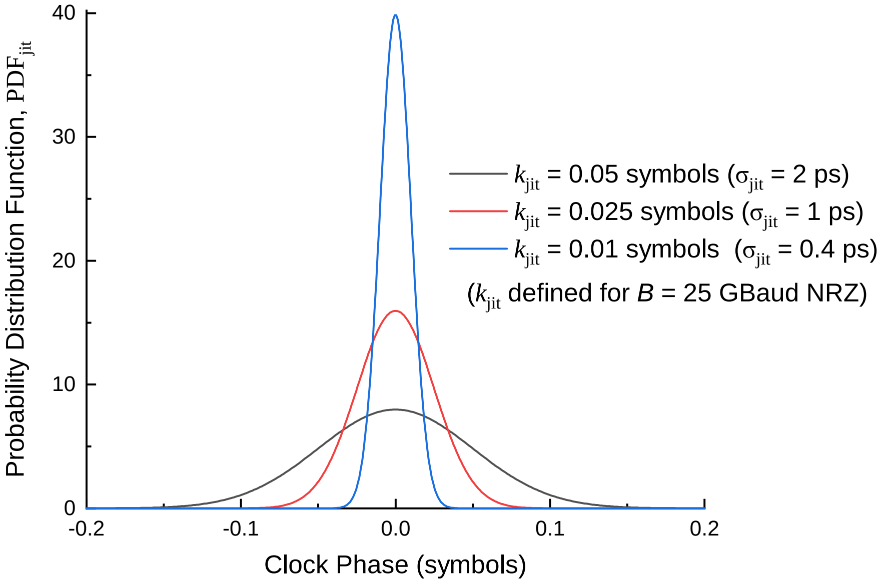

Fig. 4.16: Probability distribution function of Gaussian (random) jitter. Curves are shown for a selection of different jitter standard deviation values in terms of time and symbols at 25 GBaud.

Equation 4.74 may be substituted into the cross-correlation in Equation 4.72 to give a complete form of the integral for calculating for error probability of an NRZ-OOK signal modelled by a Gaussian impulse response, including the effect of both receiver noise and sampling clock jitter:

$$p_{e(\mathrm{post-jit})}(\phi) = \int_{-\infty}^{\infty} \frac{1}{k_{\mathrm{jit}}\sqrt{2\pi}} \exp\left(-\frac{\alpha^2}{2k_{\mathrm{jit}}^2}\right) p_e(\phi+\alpha)\, d\alpha \tag{4.75}$$

where $p_e(\phi+\alpha)$ is given by Equation 4.67, with $\phi+\alpha$ directly substituted for ϕ.

Figure 4.17 then illustrates a cross-correlation between a Gaussian jitter distribution and the error probability given in Equation 4.67 resulting from the Gaussian NRZ pulse shape. The output of the cross-correlation is the error probability when sampling at a given clock phase offset in the pulse shape after accounting for the random offset of the sampling phase position in the pulse shape by clock jitter. With a k_{jit} of 0.05 symbols, the error probability at sampling clock phase offsets other than ϕ = 0 is significantly degraded from the original error probability prior to applying jitter to the sampling point. This has the effect of reducing the acceptable range of clock phase offsets within which error probability is minimised.

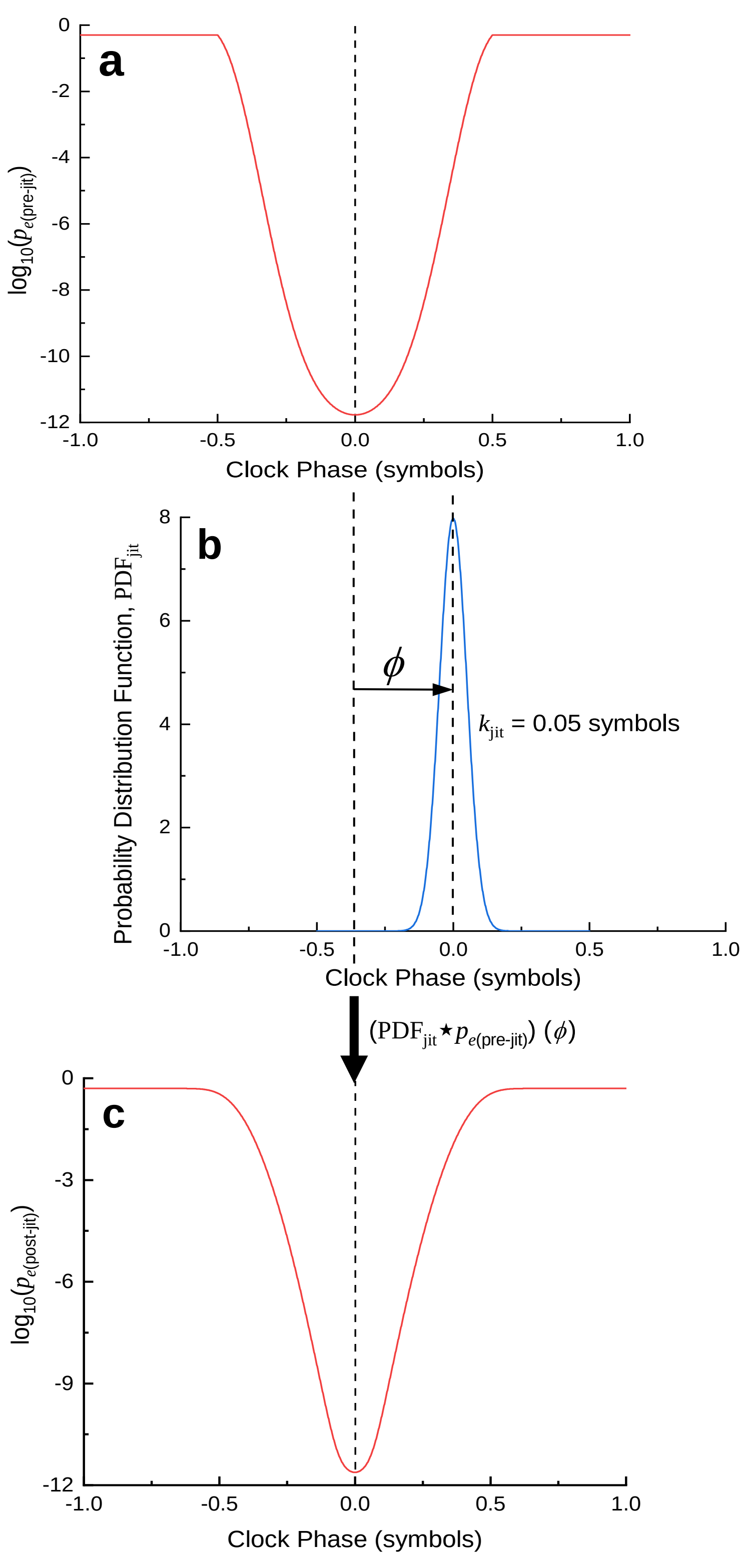

Fig. 4.17: Cross-correlation between pulse shape and jitter. **a**, the original error probability, p_e, from the pulse shape prior to applying jitter, generated in the same manner as used in Figure 4.15, for a received average optical power of -13.75 dBm. **b**, the probability density function of the jitter on the sampling phase, which is modelled by a Gaussian with an standard deviation k_{jit} of 0.05 symbols (2 ps at 25 GBaud NRZ). **c**, increased error probability, $p_{e(\text{post}-\text{jit})}$, at different sampling clock phase offsets after accounting for random sampling phase offsets occurring due to jitter.

To illustrate the effect of clock jitter on the error probability as a function of clock phase offset in a Gaussian NRZ-OOK signal, consider the -10.5 dBm error probability curve in Figure 4.15. The error probability after accounting for the jitter of the sampling clock can be calculated by performing a cross-correlation with a series of jitter standard deviations, k_{jit}. Figure 4.18 shows the outcome of this cross-correlation for the -10.5 dBm error probability curve in Figure 4.15. Error probability degradation increases with greater jitter standard deviations, k_{jit}.

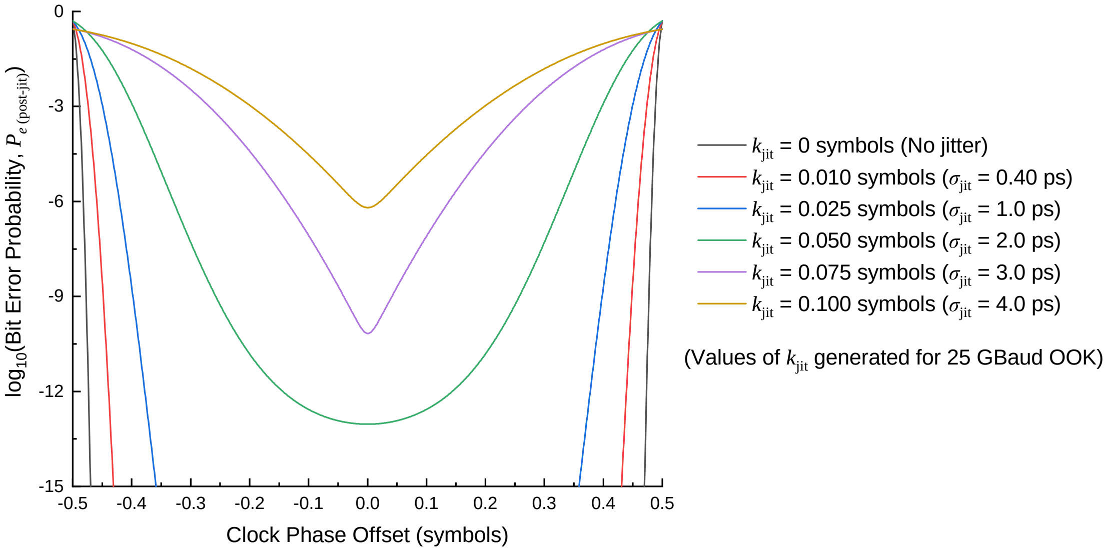

Fig. 4.18: Error probability degradation resulting from jitter of the sampling clock position. An average received optical power of -10.5 dBm was used to generate these curves is, with the black curve matching the -10.5 dBm curve in Figure 4.15. Significant eye closure occurs for jitter standard deviations beyond approximately 0.05 symbols.

To further illustrate the effect of jitter on error probability, consider the eye diagram generated in Figure 4.14b, at an average received optical power of -10.5 dBm, which is approximately equal to the minimum average received optical power for 25 GBaud OOK data centre signals [10]. Figure 4.19 shows a set of four eye diagrams at different jitter standard deviations, k_{jit}, that show the effect of applying jitter to Figure 4.14b. These were generated by evaluating the cross-correlation in Equation 4.75 on the OOK pulse shape in Equation 4.62 at different decision photocurrents, $I_D(\phi)$, using the original unjittered eye diagram in Figure 4.14b, which was generated using an average received optical power of -10.5 dBm.

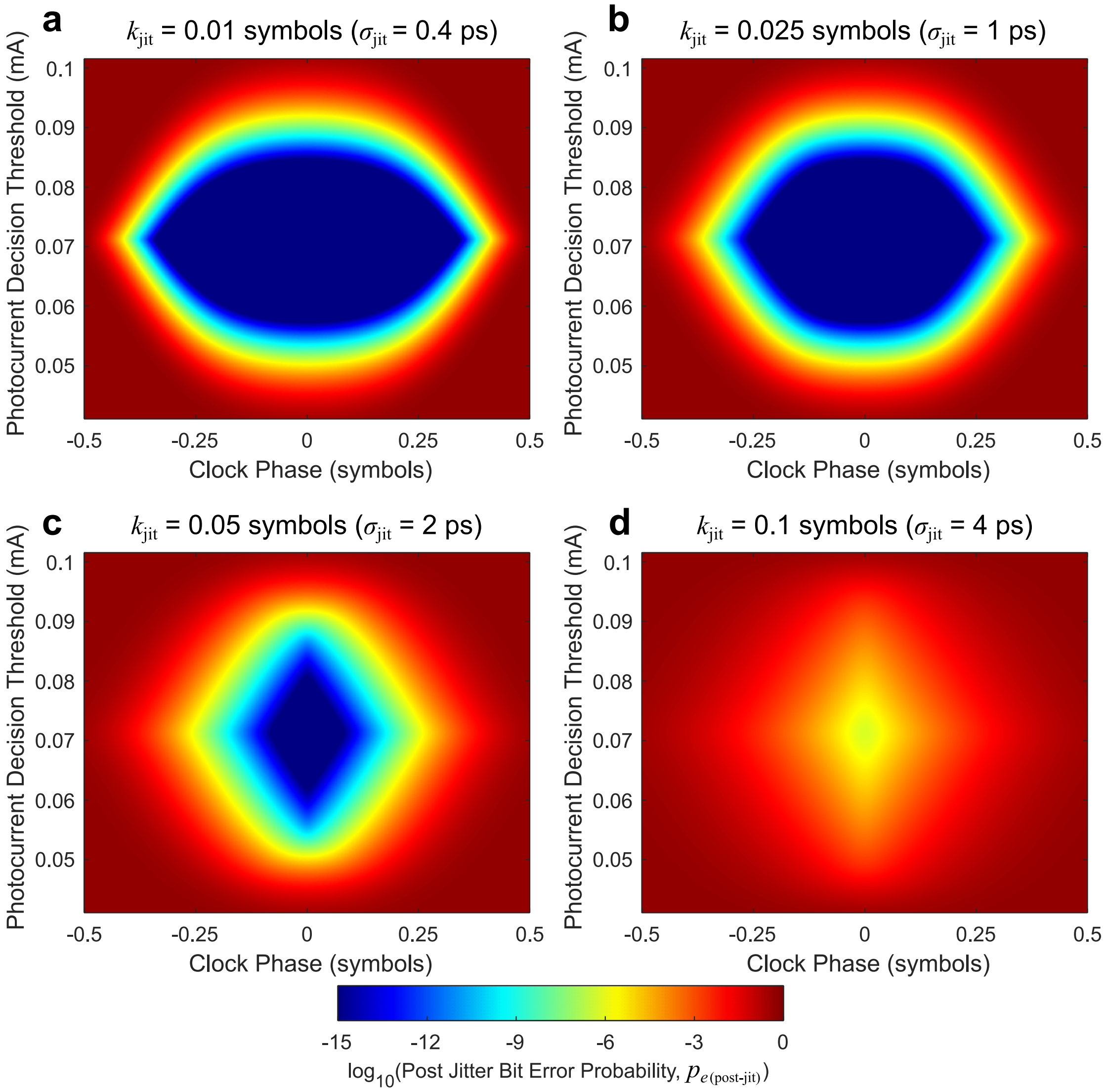

Fig. 4.19: Analytically modelled receiver eye diagrams showing error probability after applying jitter, $p_{e(\text{post}-\text{jit})}$. This is achieved using Equation 4.75 at a series of different jitter standard deviations, k_{jit}. Error probabilities smaller than 10^{-15} are displayed as a probability of 10^{-15}. All values of σ_{jit} were calculated for 25 GBaud NRZ-OOK. **a**, k_{jit} = 0.01 symbols (σ_{jit} = 0.4 ps), which results in almost no eye degradation versus the unjittered signal in Figure 4.14b; **b**, k_{jit} = 0.025 symbols (σ_{jit} = 1 ps), which results in a small reduction of the range of clock phase offsets with acceptable error probability; **c**, k_{jit} = 0.05 symbols (σ_{jit} = 2 ps), which results in a significant reduction in the range of clock phase offsets with acceptable error probability; **d**, k_{jit} = 0.10 symbols (σ_{jit} = 4 ps), where there is complete eye closure due to jitter.

4.4 Analytical Modelling of Power Penalty

The worst case clock phase offset caused by temperature variation within the data centre is greater for larger transmission distances, which results in a larger bit error probability. The magnitude of these worst-case clock phase offsets, their impact on error probability, and the resulting power penalty from using single calibration CSA-CDR without packet clock phase tracking, versus traditional asynchronous CDR, will now be assessed using the analytical model and data centre environmental conditions introduced in the previous section. These quantities will be modelled for the four main standardised data centre fibre length scales illustrated in Chapter 2 Figure 2.2 that a data centre optical switch could be used to interconnect: inter-rack ($\leq$7 m), intra-cluster ($\leq$100 m), core ($\leq$2 km) and inter-building ($\leq$10 km).

4.4.1 Effect of Clock Phase Offset (With and Without Clock Jitter)

By assuming a constant photocurrent decision threshold, I_D, of halfway between the positive and negative-going pulses, with a value of RP_{avg}, bit error probability can be calculated as a function of both clock phase offset, ϕ, and average received optical power, P_{opt}, using Equation 4.67. The result of this calculation for the photodiode characteristics in Table 4.2 is shown in Figure 4.20. The range of clock phase offsets for which bit error probability is less than 10^{-12} initially quickly increases after an average received optical power of -13 dBm, followed by an asymptotic increase in range that tends towards a width of 1 symbol beyond an average received optical power of approximately -12 dBm.

To maintain the same bit error probability at phase offsets away from the optimum phase sampling point at $\phi = 0$ symbols, an increase in average received optical power is required, which can be considered a power penalty. The power penalty can be calculated for a given error probability by following the contours in Figure 4.20. Figure 4.21 shows the power penalty incurred from operating the CDR away from the optimum sampling point, for error probabilities $< 10^{-9}$.

A small power penalty, such as 0.1 to 1 dB, might be considered a tolerable compromise if it removes the need to reacquire clock phase. However, before considering what power penalties would occur for different data centre length scales and temperature ranges, which determines the clock phase offset, the effect of clock jitter on power penalty must be included, which acts to increase the power penalty at a given clock phase offset. This impact will be considered in the next sub-section.

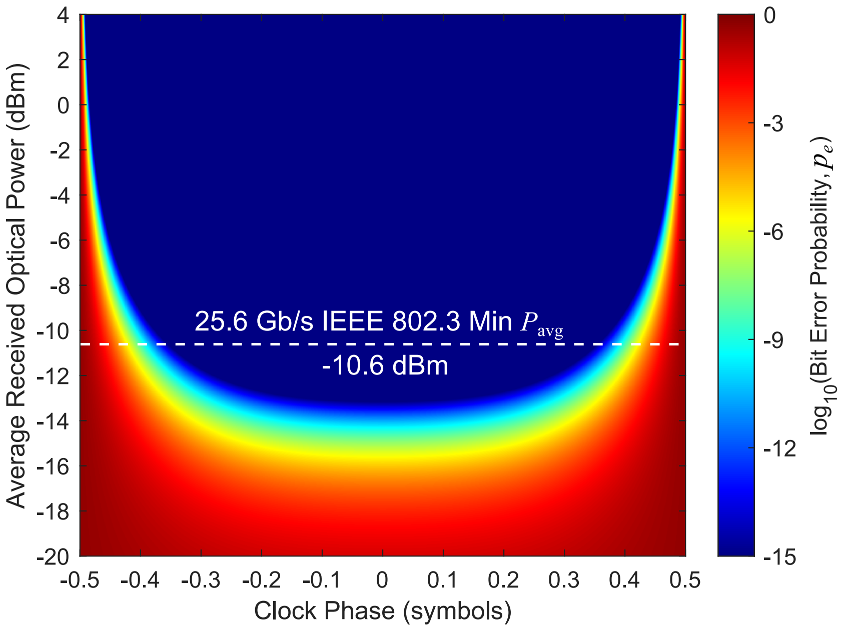

Fig. 4.20: OOK error probability, p_e, resulting from a Gaussian impulse response, calculated against clock phase offset, ϕ. k_{pls} = 0.189 was used (corresponding to a cut-off frequency f_c of 17.5 GHz for a 25 GBaud NRZ-OOK signal), and the error probability was calculated for a range of average received optical powers that includes the standardised range for data centre average received optical power of -10.6 to 4.5 dBm for 25 Gb/s NRZ-OOK [10]. This plot does not include the effect of clock jitter.

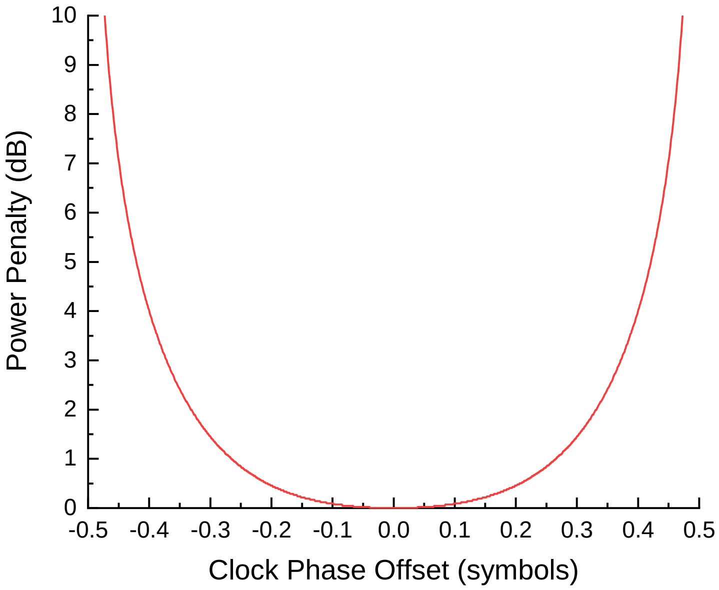

Fig. 4.21: Power penalty incurred from operating away from the optimum sampling point within the Gaussian OOK signal. The power penalty was generated using the photodiode characteristics in Table 4.2, generated from Figure 4.20 by evaluating the additional power required to achieve the same error probability as the clock phase offset is changed from the optimum sampling point. The power penalty shown here is applicable for target worst-case error probabilities of under 10^{-9}. This plot does not include the effect of clock jitter.

Figure 4.20, which gives the bit error probability as a function of clock phase and average received optical power, can be extended to include the effect of clock jitter. This can be achieved by using Equation 4.75 to cross-correlate the error probability at each average received optical power in Figure 4.20 with a Gaussian jitter probability distribution. Figure 4.23 shows the result of this cross-correlation for a series of increasing jitter standard deviations, k_{jit}. Increased jitter causes the range of clock phase offsets with an acceptable error probability, for example under 10^{-10}, to decrease. An error probability of 10^{-10} was used in this chapter to match the threshold BER used for the experimental measurements performed in Chapters 6 and 7, where further explanation of the choice of BER threshold is given.

In the same fashion as for Figure 4.20, the additional power required to maintain a set bit error probability, or power penalty, as the clock phase offset is moved away from the optimal sampling point at $\phi = 0$, can be calculated from Figures 4.23(a-d). This can be achieved by following the error probability contour for a set error probability threshold, e.g. 10^{-10}. Figure 4.22 then shows the power penalty required to maintain a set bit error probability as the clock phase offset is moved away from the optimal sampling point at $\phi = 0$. Increasing the jitter standard deviation, k_{jit}, causes a decrease in the range of clock phase offsets for which the power penalty is small, e.g. under 0.1 to 1 dB. For a jitter standard deviation of up to 0.025 ps, the decrease in the range of tolerable clock phase offsets at a given power penalty is negligible. However, for standard deviations beyond this cause increasingly large reductions in the range of tolerable clock phase offsets.

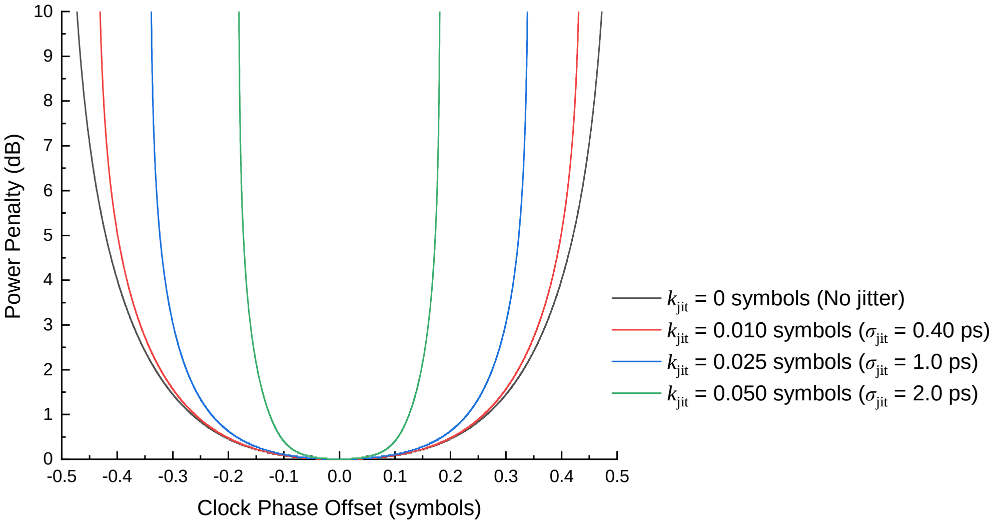

Fig. 4.22: Increase in power penalty for a Gaussian NRZ-OOK signal that includes clock jitter, versus the corresponding case without jitter shown in Figure 4.21. The power penalty at different values of the jitter standard deviation, k_{jit}, was calculated by following the 10^{-10} error probability contours against clock phase offset in Figures 4.23a-c. No power penalty is shown for Figure 4.23d as there is no region at any optical power or clock phase where the bit error probability is equal to 10^{-12}.

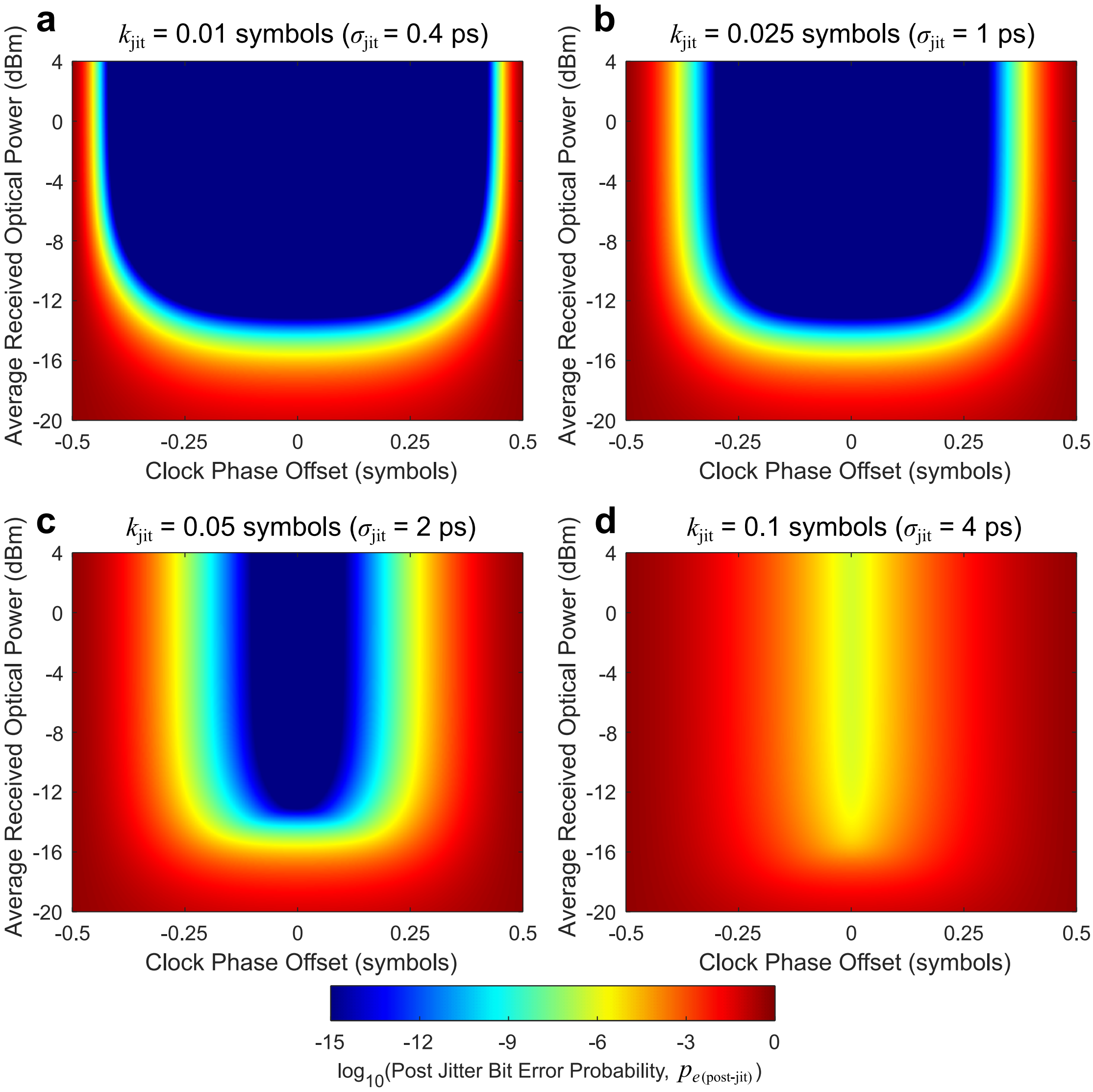

Fig. 4.23: OOK error probability, $p_{e(\text{post}-\text{jit})}$, as a function of average received optical power and clock phase offset, including the effect of jitter. This was achieved by using Equation 4.75 to cross-correlate the error probability at each average received optical power in Figure 4.20 with a Gaussian jitter probability distribution. A selection of different jitter magnitudes are plotted. Error probabilities smaller than 10^{-15} are displayed as a probability of 10^{-15}. All values of σ_{jit} were calculated for 25 GBaud NRZ-OOK. **a**, k_{jit} = 0.01 symbols (σ_{jit} = 0.4 ps), which results in a small reduction in the range of clock phase offsets with acceptable error probability; **b**, k_{jit} = 0.025 symbols (σ_{jit} = 1 ps), which results in a moderate reduction of the range of clock phase offsets with acceptable error probability; **c**, k_{jit} = 0.05 symbols (σ_{jit} = 2 ps), which results in a large reduction in the range of clock phase offsets with acceptable error probability; **d**, k_{jit} = 0.10 symbols (σ_{jit} = 4 ps), where there is complete eye closure due to jitter.

4.4.2 Effect of Temperature Variation at Different Data Centre Distances

In this subsection, the power penalty resulting from change in data centre temperature since clock phase calibration will be estimated for four different data centre distance ranges. Recall that the change in clock phase, $\Delta\phi$, due to the temperature change, ΔT, is given by Equation 4.10 to be $\Delta\phi = 2\tau LB\Delta T$, where τ is the thermal coefficient of delay of the optical fibre used to transport the clock and data, L is the distance between nodes and B is the symbol rate.

Assuming ΔT represents the change in temperature that has occurred since the clock phase was calibrated to be 0 symbols, Equation 4.10 can be substituted for the clock phase offset ϕ in Equation 4.75, to give:

$$p_{e(\mathrm{post-jit})}(\Delta T) = \int_{-\infty}^{\infty} \frac{1}{k_{\mathrm{jit}}\sqrt{2\pi}} \exp\left(-\frac{\alpha^2}{2k_{\mathrm{jit}}^2}\right) p_e(2\tau LB\Delta T + \alpha)\, d\alpha \tag{4.76}$$

where k_{jit} is the standard deviation of the jitter in symbols and $p_e(2\tau LB\Delta T + \alpha)$ is given in Equation 4.67 (with $2\tau LB\Delta T + \alpha$ substituted for ϕ), which describes the error probability resulting from the NRZ-OOK pulse shape including the effect of clock phase offset due to temperature change and jitter. $p_e(2\tau LB\Delta T + \alpha)$ is given by:

$$\begin{aligned} p_e(2\tau LB\Delta T + \alpha) = \frac{1}{4}\Bigg(&\operatorname{erfc}\left(\frac{I_1(2\tau LB\Delta T + \alpha) - RP_{\mathrm{avg}}}{\sqrt{4qI_1(2\tau LB\Delta T + \alpha)f_c + 2\sigma_{ni}^2}}\right) \\ &+ \operatorname{erfc}\left(\frac{RP_{\mathrm{avg}} - I_0(2\tau LB\Delta T + \alpha)}{\sqrt{4qI_0(2\tau LB\Delta T + \alpha)f_c + 2\sigma_{ni}^2}}\right)\Bigg) \end{aligned} \tag{4.77}$$

where q is the electronic charge of one electron, f_c is the cut-off frequency of the TIA, σ_{ni} is the input referred noise of the TIA, R is the PIN photodiode responsivity, P_{avg} is the average received optical power, and $I_1(2\tau LB\Delta T + \alpha)$ and $I_0(2\tau LB\Delta T + \alpha)$ are the positive and negative-going NRZ-OOK photocurrent pulses given in Equations 4.48 and 4.49 (where again $2\tau LB\Delta T + \alpha$ is substituted for ϕ), which are:

$$\begin{aligned} I_1(2\tau LB\Delta T + \alpha) = RP_{\mathrm{avg}}\left(\frac{r_e - 1}{r_e + 1}\right)\Bigg(&\operatorname{erf}\left(\frac{2\tau LB\Delta T + \alpha + \frac{1}{2}}{\sqrt{2}k_{\mathrm{pls}}}\right) \\ &- \operatorname{erf}\left(\frac{2\tau LB\Delta T + \alpha - \frac{1}{2}}{\sqrt{2}k_{\mathrm{pls}}}\right)\Bigg) + \frac{2RP_{\mathrm{avg}}}{r_e + 1} \end{aligned} \tag{4.78}$$

$$\begin{aligned} I_0(2\tau LB\Delta T + \alpha) = RP_{\mathrm{avg}}\left(\frac{r_e - 1}{r_e + 1}\right)\Bigg(&2 - \operatorname{erf}\left(\frac{2\tau LB\Delta T + \alpha + \frac{1}{2}}{\sqrt{2}k_{\mathrm{pls}}}\right) \\ &+ \operatorname{erf}\left(\frac{2\tau LB\Delta T + \alpha - \frac{1}{2}}{\sqrt{2}k_{\mathrm{pls}}}\right)\Bigg) + \frac{2RP_{\mathrm{avg}}}{r_e + 1} \end{aligned} \tag{4.79}$$

where r_e is the transmitter extinction ratio and k_{pls} is the standard deviation in symbols of the Gaussian filter used to obtain the NRZ-OOK photocurrent pulse shape. Finally, k_{pls} is related to the cut-off frequency of the TIA, f_c, by the symbol rate, B:

$$k_{\text{pls}} = \frac{B\sqrt{\ln 2}}{2\pi f_c} \tag{4.80}$$

The error probability after applying jitter, $p_{e(\text{post}-\text{jit})}$, can be calculated as a function of the temperature change, ΔT, the average received optical power, P_{opt} and the magnitude of the clock jitter, σ_{jit}. The power penalty as the clock phase offset induced by fibre temperature change increases can then be calculated by evaluating the increase in optical power required to maintain a set bit error probability, such as 10^{-10}.

To perform this calculation, the following parameter values will be used:

- An optical fibre thermal coefficient of delay, τ, of 42.6 ps/(km·°C), matching the measured thermal coefficient of delay of loose-tube buffered single-mode optical fibre [94], which is commonly used in data centres.

- A symbol rate, B, of 25 GBaud, a typical symbol rate used in NRZ-OOK data centre transmission [10]. While 50 GBaud PAM-4 data centre transmission is expected to be standardised in the next couple of years, and 25 GBaud PAM-4 data centre transmission is already standardised, at the time that the experimental results were acquired later in this thesis only 25 GBaud NRZ FPGA transceivers that were suitable for implementing clock phase assisted clock and data recovery were available.

- A TIA limited cut-off frequency, f_c, of 17.5 GHz, matching the -3 dB bandwidth of a typical commercial TIA for 25 Gb/s OOK reception in data centres [98].

- A TIA input referred noise, σ_{ni}, of 2 μA, matching the input referred noise of a typical commercial TIA for 25 Gb/s OOK reception in data centres [98].

- A PIN photodiode responsivity, R, of 0.7 A/W, matching the 1310/1550 nm responsivity of a typical commercial PIN photodiode designed for 25 Gb/s OOK reception in data centres [99] (note that the photodiode has a -3 dB bandwidth of 22 GHz).

- An extinction ratio, r_e, of 4 dB, matching the IEEE 803.2 standardised minimum extinction ratio for 25 GBaud NRZ-OOK data centre transmission [10].

- A sampling clock jitter standard deviation, k_{jit}, of 0.0604 symbols, matching the total jitter measured from a clock synchronised Xilinx GTY 25 GBaud transmitter (further details of this measurement are given in Chapter 6).

Figure 4.24 then shows the power penalty as a function of temperature change, at each of the four typical data centre length scales (Figure 4.24a, intra-rack (≤7 m); Figure 4.24b, intra-cluster (≤100 m); Figure 4.24c, core (≤2 km) and Figure 4.24d, inter-building (≤10 km)), for a selection of different jitter standard deviations. If a 1 dB power penalty resulting from clock phase offset from operating without the CDR is considered reasonable, then the range of acceptable temperatures for each length scale are 9.45 °C for intra-rack (≤7 m), 0.660 °C for intra-cluster (≤100 m), 0.0330 °C for core (≤2 km) and 0.0660 °C for inter-building (≤10 km).

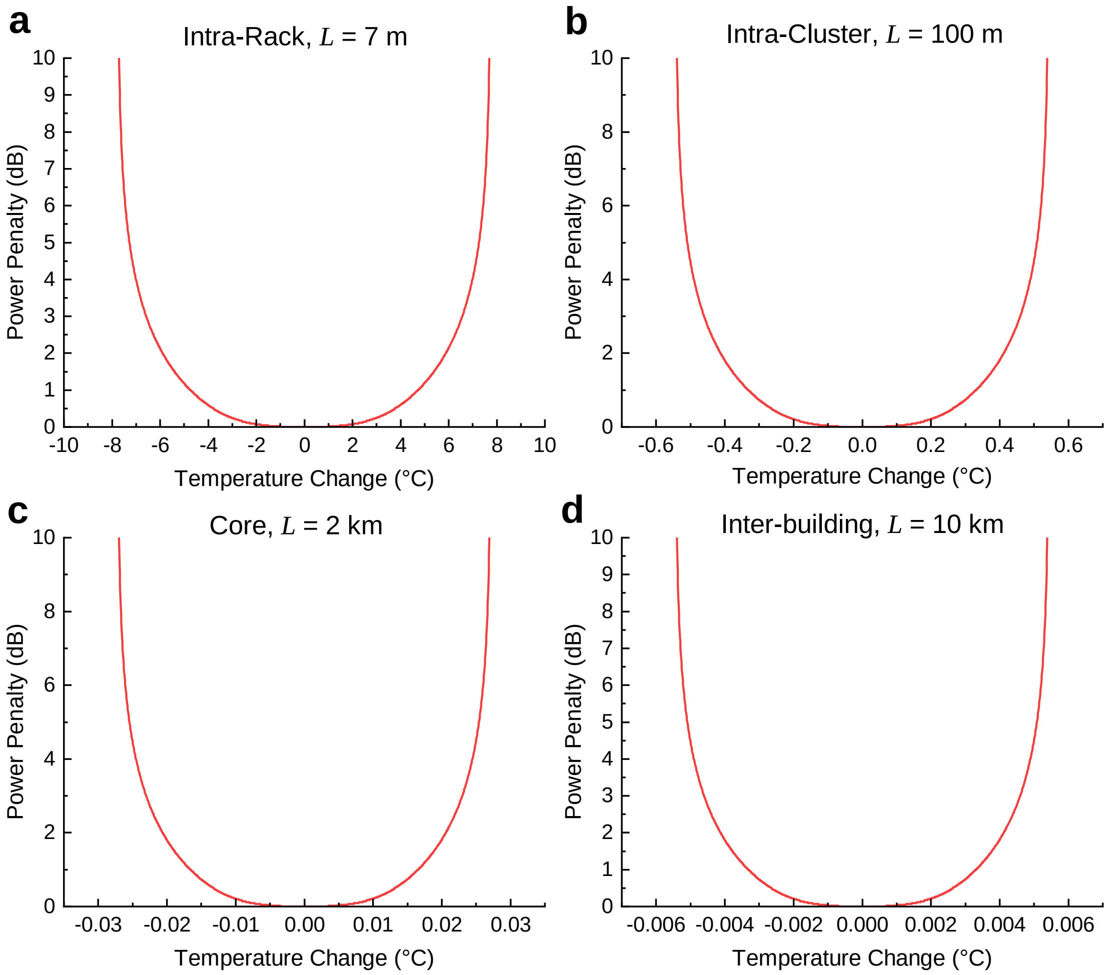

Fig. 4.24: Analytically modelled power penalty to maintain a bit error probability of 10^{-10} at a receiver, resulting from temperature change in an optically-switched network, operating with single calibration CSA-CDR without packet clock phase tracking. The power penalty was modelled at four different data centre length scales. NRZ-OOK signal reception was modelled with a symbol rate of 25 GBaud using typical data centre TIA and PIN photodiode characteristics, with a jitter standard deviation, k_{jit}, of 0.0604 symbols (based on a measurement from a synchronised Xilinx GTY 25 GBaud NRZ transmitter). **a**, intra-rack (≤7 m); **b**, intra-cluster (≤100 m); **c**, core (≤2 km); **d**, inter-building (≤10 km).

4.4.3 Summary of Analytical Modelling Assumptions

The assumptions made by the analytical modelling presented in this chapter may be summarised as:

- A transmitter, receiver, switch and clock source layout that maximises the clock phase offset due to temperature change was assumed, to assess the effect of the worst-case topology. This occurs when the transmitting node is located far from the switch, clock source and receiver, which are all co-located (as shown in Figure 4.5).

- The thermal coefficient of delay of optical fibre, τ, was modelled as linear. This is a reasonable assumption for data centre recommended operation between temperatures of 15 and 45 °C [22], as measurements of the thermal coefficient of delay of loose-tube, 250 μm tight and 900 μm semi-tight tube buffered SMF-28 show that it is strongly linear for all three types types within this recommended data centre temperature range [94].

- The NRZ pulse was modelled by a convolution between a Gaussian filter impulse response and an ideal NRZ pulse. This pulse shape was chosen in preference to other alternatives because it minimises rise and fall times while having no overshoot/undershoot [96], it allows simple, closed-form expressions to be derived to describe the pulse shape [95].

- That ISI was negligible. This was reasonable because the worst-case reduction in eye height at a -3 dB bandwidth of 17.5 GHz for 25 GBaud NRZ-OOK is 1.65% (calculated using Equations 4.18 and 4.30), assuming Gaussian NRZ pulse modelling.

- The probability of transmission of a 1 or 0 are equal, which was reasonable since encoding is used to ensure DC line balance in data centre transmission [10].

- That photodiode shot noise can be modelled by a Gaussian random process. This was reasonable as the minimum average optical power used in the analytical modelling of -20 dBm is 25 dB greater than the quantum limit of detection (about -45 dBm for a symbol rate of 25 GHz and 1550 nm transmission [86]).

- That the dark current contribution to photocurrent noise was negligible. This is reasonable as the variance of the thermal noise (modelled by the TIA input referred noise), σ_T^2, of 4×10^{-12} A^2 [98], is 250000 times greater than the variance of the dark current noise, σ_D^2, of 1.6×10^{-17} A^2, calculated using Equation 4.55 and the dark current of 9 nA of a commercial PIN photodiode used for data centre transmission [99].

- That the decision threshold was halfway between the positive-going and negative-going pulses. This is reasonable as a data centre receiver, such as those located on an FPGA, may not have the functionality to be able to adjust the decision threshold to minimise bit error probability at a given received average optical power.

- The clock jitter was modelled as a Gaussian random process. Although real-world clock jitter does consist of both deterministic and random jitter, entirely random jitter was assumed to minimise the complexity of the analytical modelling.

4.5 Discussion

As discussed in Chapter 2, worst-case data centre temperature variation of intake air for servers in a data centre rack, as given by recommendations for data centre design, may be by up to 40 °C. Server exhaust temperature may vary by up to 70 °C due to local hot-spots caused by equipment such as ToR switches. As shown in Figure 4.24, for a total jitter of 2 ps, the power penalty from single calibration CSA-CDR without packet clock phase tracking exceeds 1 dB over a 40 °C temperature range for all data centre length scales.

For intra-rack communication, single calibration CSA-CDR without packet clock phase tracking would be potentially feasible if the optical fibre interconnecting the servers and ToR switch is laid within the temperature controlled cool aisle (i.e. on the server air intake side), and the temperature range is restricted to the A1 (18 to 27 °C) class of allowable temperatures [22]. In this case, the range of acceptable temperatures for this class is under the 9.45 °C range for which a power penalty of under 1 dB results from using single calibration CSA-CDR without packet clock phase tracking. However, this is not likely to be practical. The technique relies on the use of SMF-28, which would increase cost versus the short-reach copper interconnects currently used for intra-rack communication, due to the added need for optical transceivers.

Refinement of single calibration CSA-CDR without packet clock phase tracking is required to increase the practicality of the technique. An optical fibre with a lower thermal sensitivity than SMF-28 could be used, such as HCF or multi-core fibre (MCF), to extend the temperature range over which a small power penalty results from single calibration single calibration CSA-CDR without packet clock phase tracking.

4.6 Contribution Statement

The analytical modelling presented in this chapter was conceived, constructed and numerically evaluated by K.A.C., supervised by Z.L..

Chapter 5

Single Calibration CSA-CDR
Part 2: With Packet Clock Phase Tracking

5.1 Introduction

An alternative approach to performing single calibration CSA-CDR is to allow the receiver CDRs to track the clock phase of incoming data packets. The process of CDR locking to incoming packets is simplified by having established clock frequency and phase synchronisation. Under frequency synchronisation, CDR reduces to a phase extraction process. Furthermore, phase extraction is simplified by the initial single calibration synchronisation step that ensures that whenever a new incoming packet is received, the initial sampling phase used by the CDR circuit is close to the ideal locked value at $\phi = 0$, at the centre of the eye. The key performance metric in such a system is CDR locking time, defined here as the time taken for the bit error probability to decay below some threshold value.

This chapter will extend the analysis in Chapter 4 to model CDR locking time for single calibration CSA-CDR with packet clock phase tracking. The analysis will assume the use of a digital phase interpolator CDR circuit, due to the wide commercial prevalence of this CDR family due to its stability, small silicon area and low power consumption [69]. This analysis will consist of a more detailed exploration of the CDR locking process for nodes connected to an optical switch, a mathematical description of the behaviour of digital phase interpolator CDRs, followed by an extension of the analytical modelling in the previous Chapter, and finally CDR locking time in optically-switched data centre networks with single calibration CSA-CDR with packet clock phase tracking at different data centre length scales will be estimated. The analytical modelling is presented for the first time in this thesis.

5.2 CDR Locking Process in a Burst-Mode Receiver

Consider a single receiver connected to N transmitters through an $N \times N$ optical switch, as illustrated in Chapter 3 Figure 3.8. At this single receiver, data packets will be arriving from N different transmitters. Each transmitter to receiver connection forms a pair, each with its own unique clock phase offset. If distributed frequency synchronisation is used such that all transmitters and receivers have the same clock frequency, then these unique clock phase offsets only change due to temperature and jitter, in the same fashion as discussed in Chapter 4.

However, if the CDR circuits within receivers are allowed to lock to the clock embedded within incoming packets, then, upon arrival of a new packet from a transmitter, a receiver phase interpolator CDR will shift its phase to match the embedded clock within the data packet. This is in contrast to the single calibration CSA-CDR without packet clock phase tracking approach explored in Chapter 4, where the CDR circuit in each receiver is only used to calibrate the initial transmission phase values. Figure 5.1 illustrates a simple case where packets arrive at a receiver, Rx, from two different transmitters, Tx0 and Tx1. Each of the two transmitter to receiver pairs, Tx0→Rx and Tx1→Rx, is associated with its own unique clock phase offset, $\phi_{\text{Tx0}\to\text{Rx}}$ and $\phi_{\text{Tx1}\to\text{Rx}}$ respectively.

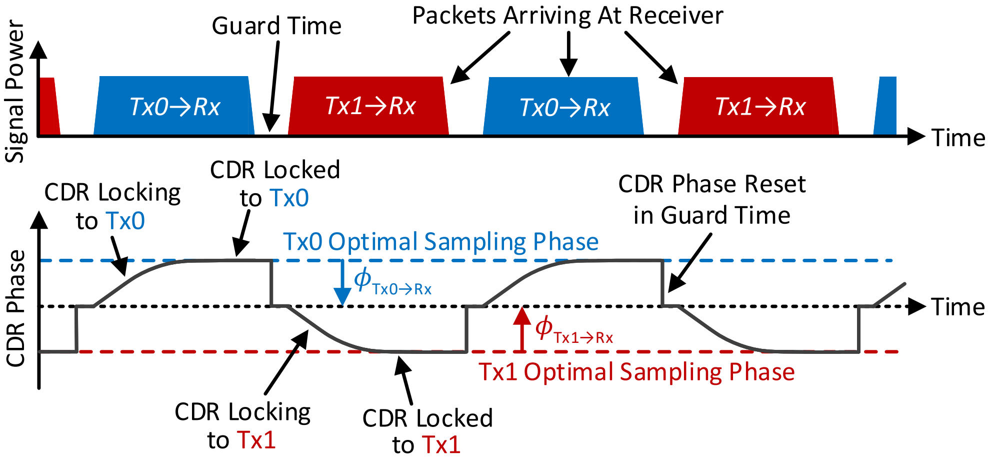
Guard Time
Packets Arriving At Receiver
Signal Power
Tx0→Rx
Tx1→Rx
Tx0→Rx
Tx1→Rx
Time
CDR Locking to Tx0
CDR Locked to Tx0
Tx0 Optimal Sampling Phase
CDR Phase Reset in Guard Time
CDR Phase
$\phi_{\text{Tx0}\to\text{Rx}}$
$\phi_{\text{Tx1}\to\text{Rx}}$
Time
Tx1 Optimal Sampling Phase
CDR Locking to Tx1
CDR Locked to Tx1

Fig. 5.1: CDR locking to incoming data packets arriving at a receiver, Rx, from two different transmitters, Tx0 and Tx1. For each incoming packet, an initial period of CDR locking occurs while the CDR adjusts its CDR phase to close to the optimal sampling phase for that transmitter to receiver pair, after which the CDR is locked. Each transmitter to receiver pair is associated with its own unique clock phase offset, in this case $\phi_{\text{Tx0}\to\text{Rx}}$ and $\phi_{\text{Tx1}\to\text{Rx}}$. These unique clock phase offsets change with phase shift due to temperature and jitter. The CDR phase is reset to a constant value within the guard band between packets to minimise the magnitude of clock phase shift required between packets.

5.3 Link Performance Metrics

In a burst-mode optically switched-system, where packets arrive from different transmitters, and clock phase lock is acquired on a packet-by-packet basis, two metrics of link performance for each transmitter to receiver pair could feasibly be used:

1. **Overall Packet Bit Error Probability:** the overall bit error probability for packets arriving at a given receiver that originate from a specific transmitter, which is obtained by measuring bit error rate across many arriving packets over a long time interval.

2. **Clock and Data Recovery Locking Time:** the time taken for bit error probability to decay below a threshold probability value (such as 10^{-10}) as the clock and data recovery circuit moves the sampling phase towards the optimum sampling point, which is obtained by measuring bit error rate at each time point in arriving packets from a specific transmitter over a long time interval.

These two quantities can be defined for packets arriving at a receiver from a specific transmitter:

- If $p_e(t)$ is the bit error probability as a function of time, t, since the beginning of packet reception at $t = 0$ and $t = t_{\text{pkt-len}}$ is the time at which the reception of each packet completes (i.e. $t_{\text{pkt-len}}$ is the time taken to receive the data packet), then the overall bit error probability, $p_{e(\text{overall})}$, may be obtained using:

$$p_{e(\text{overall})} \triangleq \frac{1}{t_{\text{pkt-len}}} \int_0^{t_{\text{pkt-len}}} p_e(t)dt \tag{5.1}$$

- The CDR locking time, t_{lock}, can be defined as the time, t, since the beginning of packet reception at $t = 0$, that bit error probability first falls below a threshold error probability of $p_{e(\text{lock})}$.

This thesis will use CDR locking time as the main metric for evaluating link performance in optical switches with CSA-CDR. This decision was made because overall bit error probability is dependent on packet length, which may be variable in data centre optical switches. Longer packet length would act to decrease overall bit error probability. CDR locking time gives a clearer indication of CDR behaviour as it indicates the point at which bit error probability has decreased below a reasonable threshold, and is unaffected by packet length. Nonetheless, for some later figures, BER will be given as a secondary metric. In this chapter, a threshold error probability, $p_{e(\text{lock})}$, of 10^{-10} will be used, to match the experimental measurements performed in Chapters 6 and 7.

5.4 Analytical Modelling of the Evolution of Bit Error Probability during CDR Locking

This section will analytically model bit error probability during the CDR locking process shown in Figure 5.1, by extending the analytical modelling used in the previous Chapter, assuming the use of a digital phase interpolator CDR.

5.4.1 Clock Phase Error of a Phase Interpolator CDR Circuit

The phase measurement functionality in a digital phase interpolator CDR is performed by a bang-bang phase detector (BB-PD). This component acts as an Alexander phase detector, which outputs a phase error of +1 when a clock edge arrive before the data edge, and outputs a phase error of -1 when a clock edge arrives after the data edge. The clock waveform is periodic, with period $1/B$ (or $1/(2B)$ if both the rising and falling edges of the clock are used to sample the data), and so this phase error behaviour is also periodic. As illustrated by the blue line in Figure 5.2, this can be mathematically defined as a repeating step function [102]:

$$\Phi_{\text{ideal}}(\phi) \triangleq \begin{cases} -1, & \text{if } n - \frac{1}{2} < \phi < n \\ 0, & \text{if } \phi = \frac{1}{2}n \\ +1, & \text{if } n < \phi < n + \frac{1}{2} \end{cases} \tag{5.2}$$

where $n \in \mathbb{Z}$

In practise, when the phase error is averaged over a large number of incoming data edges (for example, 64 edges), the phase error function is linearised by clock jitter of the data edge phase, as illustrated by the red line in Figure 5.2. This linearisation may be calculated by performing a convolution between the jitter PDF and the step function in Equation 5.2, if the jitter is assumed to be Gaussian distributed as defined in Equation 4.74 and is not large in magnitude ($k_{\text{jit}} < 1/(8\sqrt{2})$), which results in a final phase error of approximately (see Appendix A for a full derivation):

$$\mathrm{E}[\Phi_{\text{jittered}}(\phi)] \approx \begin{cases} \operatorname{erf}\left(\frac{\phi-n}{\sqrt{2}k_{\text{jit}}}\right), & \text{if } n - \frac{1}{4} < \phi \leq n + \frac{1}{4} \\ -\operatorname{erf}\left(\frac{\phi-n-\frac{1}{2}}{\sqrt{2}k_{\text{jit}}}\right), & \text{if } n + \frac{1}{4} < \phi \leq n + \frac{3}{4} \end{cases} \tag{5.3}$$

where $n \in \mathbb{Z}$ and $k_{\text{jit}} < \frac{1}{8\sqrt{2}}$

Equation 5.3 consists of two sets of terms, even and odd, each associated with the two different types of linear phase error regions. Figure 5.2 illustrates the centre-most even term and the two nearest odd terms from Equation 5.3.

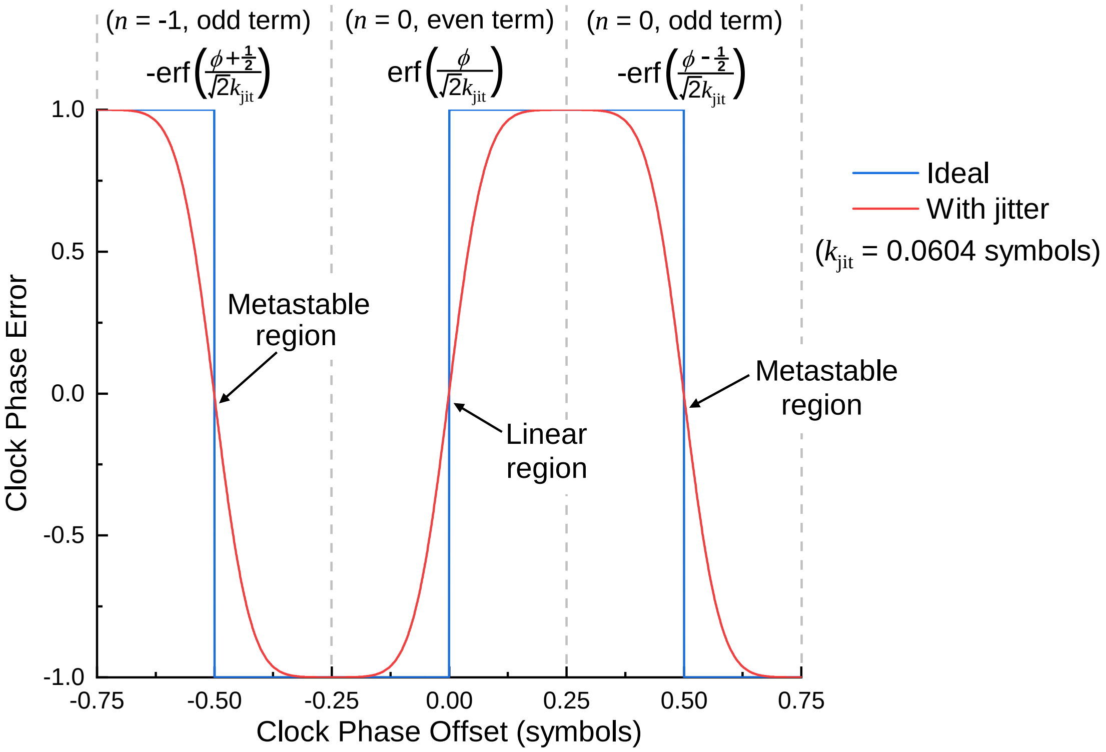

Fig. 5.2: Bang-bang phase detector phase error. The phase error of an ideal bang-bang phase detector is shown (blue line), as well as the expectation of the phase error of a bang-bang phase detector in the presence of jitter between the incoming data and the sampling clock (red line). Firstly, the jitter acts to create a linear phase error region in the central region in the vicinity of 0 symbols, which is repeated at multiples of n symbols corresponding to the periodicity of the sampling clock. Secondly, the jitter also acts to create metastable regions at multiples of $n - \frac{1}{2}$ symbols, which lead to long CDR locking time if the initial sampling phase upon data packet arrival begins in these regions due to slow movement of the CDR phase. A value of k_{jit} of 0.0604 symbols was used to generate the two jittered phase errors as an example.

The first type of linear region, corresponding to even terms in Equation 5.3, is centred on integer multiples of a symbol period, including $\phi = 0$. In this region, the clock phase is close to the ideal sampling clock phase offset. The phase error in this region is mathematically described by the positive error function in Equation 5.3. Physically, the linearisation of the clock phase error in this region acts to slow the rate of CDR phase movement as the CDR approaches the ideal sampling phase at $\phi = 0$.

The second type of linear region, corresponding to odd terms in Equation 5.3, is centred on clock phase offsets at every other half symbol period. This is the metastable region discussed in Chapter 2, which is a serious limitation of burst-mode phase interpolator CDRs preventing sub-nanosecond CDR locking time. This due to long CDR locking times that occur when the initial sampling phase begins in this region due to slow movement of the CDR phase. The phase error in this region is mathematically described by the negative error function in Equation 5.3.

5.4.2 Evolution of Clock Phase Offset within Received Data Packets

The behaviour of the CDR phase within a stream of packets arriving from a single transmitter will now be analytically modelled. In each packet, the CDR phase is initially offset by ϕ_0 from the ideal sampling phase at ϕ = 0. This initial clock phase offset is due to a combination of fibre temperature change that has occurred in the data centre since clock phase calibration, in addition to an offset due to clock jitter. The time elapsed during the reception of each packet is defined as t, where t = 0 is the time at the beginning of the packet and t = $t_{\mathrm{pkt-len}}$ is the time at which the reception of each packet completes. In an ideal case without jitter, the movement of the CDR phase, $d\phi$, at a clock phase offset, ϕ, determined over a time interval, dt, is given by the clock phase error measured by an ideal bang-bang phase detector, $\Phi_{\mathrm{ideal}}(\phi)$, multiplied by the proportional gain constant of the phase interpolator CDR circuit, p:

$$d\phi = -p\,\Phi_{\mathrm{ideal}}(\phi)dt \tag{5.4}$$

In practise, the impact of clock jitter needs to be accounted for, and so the expectation of the movement of the CDR phase, $d\phi$, at a clock phase offset, ϕ, determined over a time interval, dt, is given by the expectation of the clock phase error measured by the bang-bang phase detector, $\mathrm{E}[\Phi_{\mathrm{jittered}}(\phi)]$, multiplied by the proportional gain constant of the phase interpolator CDR circuit, p:

$$\mathrm{E}[d\phi] = -p\,\mathrm{E}[\Phi_{\mathrm{jittered}}(\phi)]dt \tag{5.5}$$

where $k_{\mathrm{jit}} < \frac{1}{8\sqrt{2}}$.

If Gaussian distributed jitter is assumed, Equation 5.3 can be substituted into Equation 5.5 to obtain the expectation of the small movement of phase, $\mathrm{E}[d\phi]$, in the presence of Gaussian distributed jitter:

$$\mathrm{E}[d\phi] \approx \begin{cases} -p\,\mathrm{erf}\left(\frac{\phi-n}{\sqrt{2}k_{\mathrm{jit}}}\right)dt, & \text{if } n-\frac{1}{4} < \phi \leq n+\frac{1}{4} \\ p\,\mathrm{erf}\left(\frac{\phi-n-\frac{1}{2}}{\sqrt{2}k_{\mathrm{jit}}}\right)dt, & \text{if } n+\frac{1}{4} < \phi \leq n+\frac{3}{4} \end{cases} \tag{5.6}$$

where $n \in \mathbb{Z}$ and $k_{\mathrm{jit}} < \frac{1}{8\sqrt{2}}$.

The expectation of the clock phase offset, E$[\phi(t)]$, the average clock phase offset that observed as a function of time, t, since the start of reception of each packet at t = 0, could ideally be obtained by finding the solution to Equation 5.6 rearranged as a series of differential equations corresponding to each odd and each even term, with the differential equation corresponding to the centre-most even term (with n = 0) being:

$$\int \frac{1}{\operatorname{erf}\left(\frac{\phi}{\sqrt{2}k_{\text{jit}}}\right)} d\phi = -\int p dt \tag{5.7}$$

However, as the left hand-side integral in Equation 5.7 has no known analytical solution, alternative approaches are required to obtain E$[\phi(t)]$. There are two potential avenues for evaluating the expectation of the clock phase offset, E$[\phi(t)]$, as a function of time, t, within packets: *1)* Use an approximation to Equation 5.6, such as treating the rate-of-change of the expectation of the clock phase error, $\Phi_{\text{jittered}}(\phi)$, as equal to the rate-of-change of Equation 5.6 at ϕ = n and ϕ = n+$\frac{1}{2}$, and then use this approximation to obtain an approximate function for E$[\phi(t)]$. *2)* Evaluate E$[\phi(t)]$ numerically, without obtaining an equation for E$[\phi(t)]$, replicating the behaviour of a CDR mathematically.

Figure 5.3 compares a linear approximation of the expectation of the clock phase error in Equation 5.6, to the exact expectation of the clock phase error in Equation 5.6 and to the ideal clock phase error, for a bang-bang phase detector in the presence of Gaussian jitter with a standard deviation of k_{jit} = 0.0604 symbols. The linear approximation is a close fit to the exact clock phase error for clock phase offsets in the vicinity of ϕ = 0, such as $|\phi| <$ 0.025 symbols for k_{jit} = 0.0604 symbols. It is therefore a good approximation for non-burst mode applications where the CDR remains always locked after first activating a point-to-point link.

However, a linear approximation deviates significantly from the exact clock phase error in the presence of jitter at intermediate clock phase offsets, such as 0.05 symbols $<$ $\phi <$ 0.1 symbols for k_{jit} = 0.0604 symbols, as can be seen in Figure 5.3. Consequently, a linear approximation would cause the CDR locking time to be underestimated, as the CDR movement rate would be estimated to be significantly greater than it would be for an exact numerical calculation. Due to the underestimation of the CDR locking time that would result from using a linear approximation for E$[\Phi_{\text{jittered}}(t)]$, E$[\phi(t)]$ will be obtained numerically rather than using a linear approximation. The method by which this will be performed will now be described.

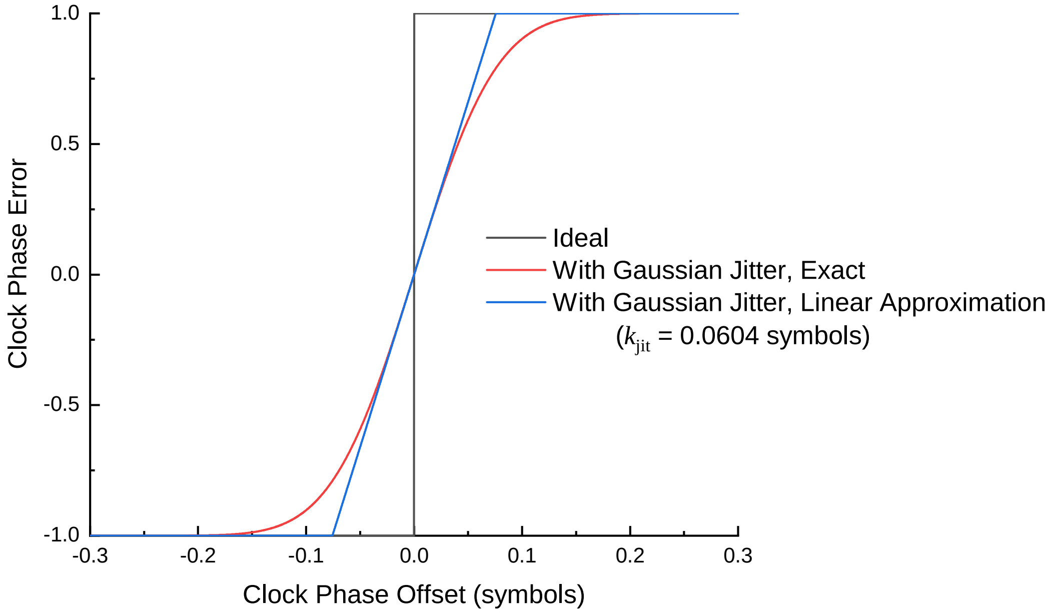

Fig. 5.3: Comparison between the linear approximation to clock phase error and a calculation including jitter. Clock phase error measured by an ideal bang-bang phase detector, the expectation of the clock phase error measured by a bang-bang phase detector in the presence of Gaussian distributed jitter, and the linear approximation of the clock phase error of a bang-bang phase detector in the presence of Gaussian distributed jitter. A value of k_{jit} of 0.0604 symbols was used to generate the two jittered clock phase errors.

Consider the operation of a digital phase interpolator CDR. At some time point since the beginning of arrival of a data packet, t, the CDR measures the clock phase offset, $\phi(t)$, over a measurement time interval, dt. The average clock phase offset measured by the CDR will be at at the expectation of the clock phase offset, $\mathrm{E}[\phi(t)]$. No clock phase shift is applied to the CDR sampling clock during the small time interval, dt, so the expectation of the clock phase offset throughout the entire time interval may be assumed to be constant, with magnitude $\mathrm{E}[\phi(t)]$. After the measurement interval has ended, the CDR then applies a small clock phase shift to its sampling clock, the expectation of which is $\mathrm{E}[d\phi(t)]$. The new clock phase offset, after the small clock phase shift is applied at the end of the measurement time interval, is then $\mathrm{E}[\phi(t + dt)]$. This behaviour may be written mathematically as:

$$\mathrm{E}[\phi(t + dt)] = \mathrm{E}[\phi(t)] + \mathrm{E}[d\phi(t)] \tag{5.8}$$

Equation 5.6, which gives the expectation of the clock phase shift, E$[d\phi(t)]$, as a function of the expectation of the clock phase offset, E$[\phi(t)]$, in the presence of Gaussian distributed jitter, may then be substituted into Equation 5.8. This substitution gives a full expression that can be used to numerically obtain the expectation of the clock phase offset as a function of time, E$[d\phi(t)]$, given that the clock phase offset at the beginning of the packet at t = 0, is ϕ_0. The full expression is:

$$\mathrm{E}[\phi(t+dt)] \approx \begin{cases} \mathrm{E}[\phi(t)] - p\,\mathrm{erf}\left(\frac{\mathrm{E}[\phi(t)]-n}{\sqrt{2}k_{\mathrm{jit}}}\right)dt, & \text{if } n-\frac{1}{4} < \mathrm{E}[\phi(t)] \leq n+\frac{1}{4} \\ \mathrm{E}[\phi(t)] + p\,\mathrm{erf}\left(\frac{\mathrm{E}[\phi(t)]-n-\frac{1}{2}}{\sqrt{2}k_{\mathrm{jit}}}\right)dt, & \text{if } n+\frac{1}{4} < \mathrm{E}[\phi(t)] \leq n+\frac{3}{4} \end{cases} \tag{5.9}$$

where $n \in \mathbb{Z}$, $k_{\mathrm{jit}} < \frac{1}{8\sqrt{2}}$ and $\mathrm{E}[\phi(0)] = \phi_0$.

The clock phase offset can then be evaluated as a function of time for a typical 25 Gb/s CDR. For this purpose, the following parameter values will be assumed:

- A Gaussian jitter standard deviation, k_{jit}, of 0.0604 symbols, measured using a digital communications analyser (see Chapter 6 for measurement of this parameter).
- A CDR proportional gain, p, of 5.77×10^6 symbols/s, measured from the CDR of a Xilinx GTY 25 GBaud NRZ transceiver by recording the rate of change of sampling phase when the incoming data signal was toggled between a clock phase offset of 0 symbols and $\frac{15}{64}$ symbols.
- A measurement time interval, dt, of 2.56 ns, based on the output rate of clock phase offset values from a Xilinx 25 Gb/s GTY transceiver CDR of once every 64 symbols.

Figure 5.4 then shows the expectation of the clock phase offset as a function of time, E$[\phi(t)]$, evaluated using Equation 5.9, for initial clock phase offsets ranging from -0.5 to 0.5 symbols. Outside of the linear region of the bang-bang phase detector where jitter decreases the magnitude of the clock phase error, the rate-of-decrease of the clock phase offsets are linear. Within the linear region of the bang-bang phase detector, jitter decreases the magnitude of the clock phase error and slows the rate-of-decrease of the clock phase offsets. The blue and red stepped curves correspond to the CDR locking behaviour shown for Tx0 and Tx1 in Figure 5.3 respectively. The change of the clock phase offsets as a function of time for the positive and negative initial clock phase offsets of equal absolute magnitude are reflected through ϕ = 0 because the Gaussian jitter distribution is symmetric. The initial rate-of-change of clock phase is slow for values of ϕ_0 that are close to $|\phi|$ = 0.5 due to metastability of the CDR in this clock phase offset region. At exactly $|\phi|$ = 0.5, the CDR clock phase does not move at all.

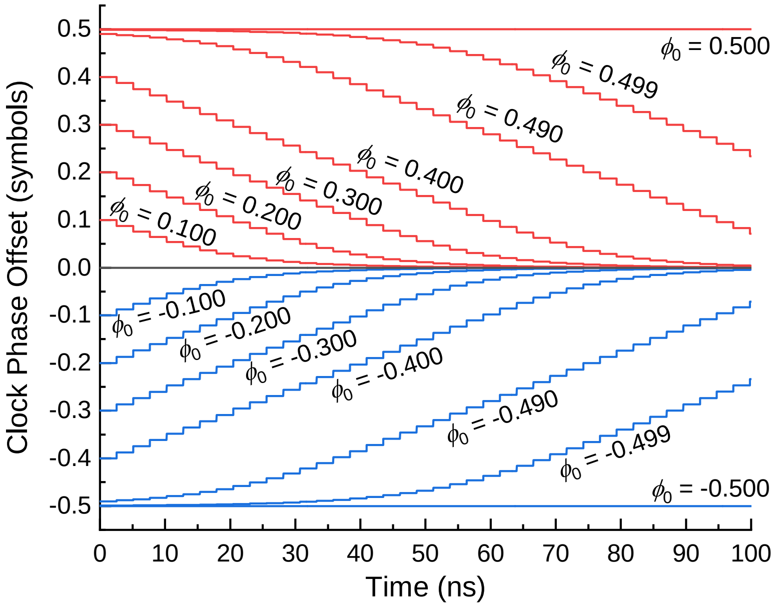

Fig. 5.4: Clock phase offset as a function of time since the beginning of packet reception, obtained numerically using Equation 5.9. A series of different values of the initial starting phase, ϕ_0, are used, from -0.5 to 0.5 symbols. These could correspond to clock phase offsets arising from temperature change, or from clock jitter. Values of ϕ_0 close to ±0.5 symbols cause long CDR locking times due to metastability of the clock phase sampling position. The rate of change of the sampling phase also slows as the clock phase offset approaches the ideal sampling position at 0 symbols. A Gaussian jitter standard deviation of k_{jit} of 0.0604 symbols, a CDR proportional gain, p, of 5.3×10^{6} symbols Hz, and a measurement time interval, dt, of 2.56 ns, were used to generate the four clock phase offsets as a function of time.

5.4.3 Evolution of Bit Error Probability within Received Data Packets (Not Including the Impact of Jitter on Clock Phase)

This subsection will obtain equations describing bit error probability as a function of time since first packet reception, using the equations describing the clock phase offset as a function of time since first packet reception, which were derived in the previous subsection, and using the equations describing bit error probability as a function of pulse shape, which were derived in Chapter 4. These derived equations will describe the evolution of bit error probability against time since first packet reception. They will model the impact of receiver noise, limited receiver bandwidth, limited extinction ratio, average received optical power and the impact of jitter on the CDR phase error, and hence on the rate-of-change of clock phase offset.

The expectation of the phase offset as a function of time, $\mathrm{E}[\phi(t)]$, given in Equation 5.9, can be substituted for ϕ in Equations 4.48 and 4.49, which describe the photocurrent pulse shapes as a function of clock phase. These are in turn substituted into Equation 4.62, to give the error probability as a function of time since the beginning of packet reception, $p_e(t)$:

$$p_e(t) = \frac{1}{4}\left(\operatorname{erfc}\left(\frac{I_1(t) - RP_{\text{avg}}}{\sqrt{4qI_1(t)f_c + 2\sigma_{ni}^2}}\right) + \operatorname{erfc}\left(\frac{RP_{\text{avg}} - I_0(t)}{\sqrt{4qI_0(t)f_c + 2\sigma_{ni}^2}}\right)\right) \tag{5.10}$$

where q is the electronic charge of one electron, f_c is the cut-off frequency of the TIA, σ_{ni} is the input referred noise of the TIA, and $I_1(t)$ and $I_0(t)$ are the magnitude of the positive and negative NRZ-OOK photocurrents given in Equations 4.48 and 4.49 as a function of time since the beginning of packet reception:

$$\begin{aligned} I_1(t) = RP_{\text{avg}}\left(\frac{r_e - 1}{r_e + 1}\right)&\left(\operatorname{erf}\left(\frac{\mathrm{E}[\phi(t)] + \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right) - \operatorname{erf}\left(\frac{\mathrm{E}[\phi(t)] - \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)\right) \\ &+ \frac{2RP_{\text{avg}}}{r_e + 1} \end{aligned} \tag{5.11}$$

$$\begin{aligned} I_0(t) = RP_{\text{avg}}\left(\frac{r_e - 1}{r_e + 1}\right)&\left(2 - \operatorname{erf}\left(\frac{\mathrm{E}[\phi(t)] + \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right) + \operatorname{erf}\left(\frac{\mathrm{E}[\phi(t)] - \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)\right) \\ &+ \frac{2RP_{\text{avg}}}{r_e + 1} \end{aligned} \tag{5.12}$$

where r_e is the transmitter extinction ratio, k_{pls} is the standard deviation in symbols of the Gaussian filter used to obtain the NRZ-OOK photocurrent pulse shape, R is the PIN photodiode responsivity and P_{avg} is the average received power. The expectation of the clock phase offset as a function of time since the beginning of packet reception, $\mathrm{E}[\phi(t)]$, is then evaluated numerically using Equation 5.9.

Lastly, k_{pls} is related to the cut-off frequency of the TIA, f_c, by the symbol rate, B:

$$k_{\text{pls}} = \frac{B\sqrt{\ln 2}}{2\pi f_c} \tag{5.13}$$

Figure 5.5 shows the bit error probability as a function of time since packet reception, evaluated using the same set of parameters and assumptions used in Chapter 4 Sub-Section 4.4.2, for a series of five different initial clock phase offset magnitudes, $|\phi_0|$. The bit error probability plots for a positive and negative initial clock phase of the same magnitude are equal because the pulse and jitter are both symmetric about $\phi = 0$. The average received optical power, P_{opt}, was been chosen to be -10.5 dBm, matching a typical intra-data centre standardised minimum average received optical power for 25 GBaud OOK reception [10].

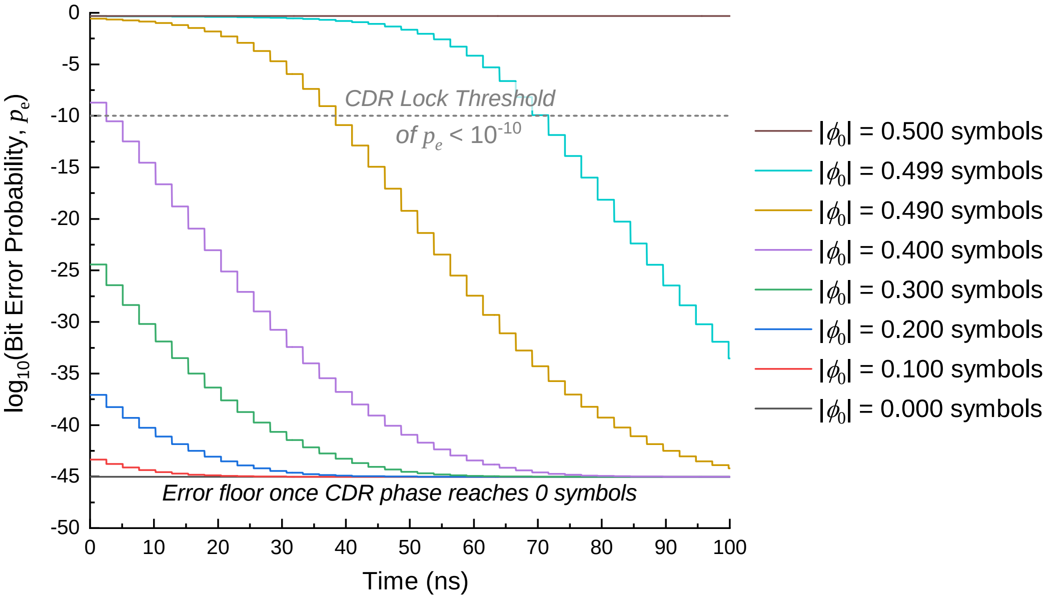

Fig. 5.5: Bit error probability as a function of time since the beginning of packet reception. The bit error probability was obtained using Equations 5.10, 5.11 and 5.12, using the initial clock phase offset magnitudes, $|\phi_0|$, from Figure 5.4. The bit error probability plots for a positive and negative initial clock phase of the same magnitude are equal because the pulse and jitter are both symmetric about $\phi = 0$. The average received optical power, P_{opt}, was -10.5 dBm, matching a typical intra-data centre standardised minimum average received optical power for 25 GBaud OOK reception [10]. A CDR locking time threshold error probability of 10^{-10} is also shown.

Bit error probability can also be plotted against initial clock phase offset, ϕ_0, at different time intervals, t, since the beginning of packet reception. This is shown in Figure 5.6. Moving vertically from one coloured line to the nearest adjacent line represents an increase of t of 2.56 ns, and so the figure illustrates the CDR locking process at each initial clock phase offset with fine granularity.

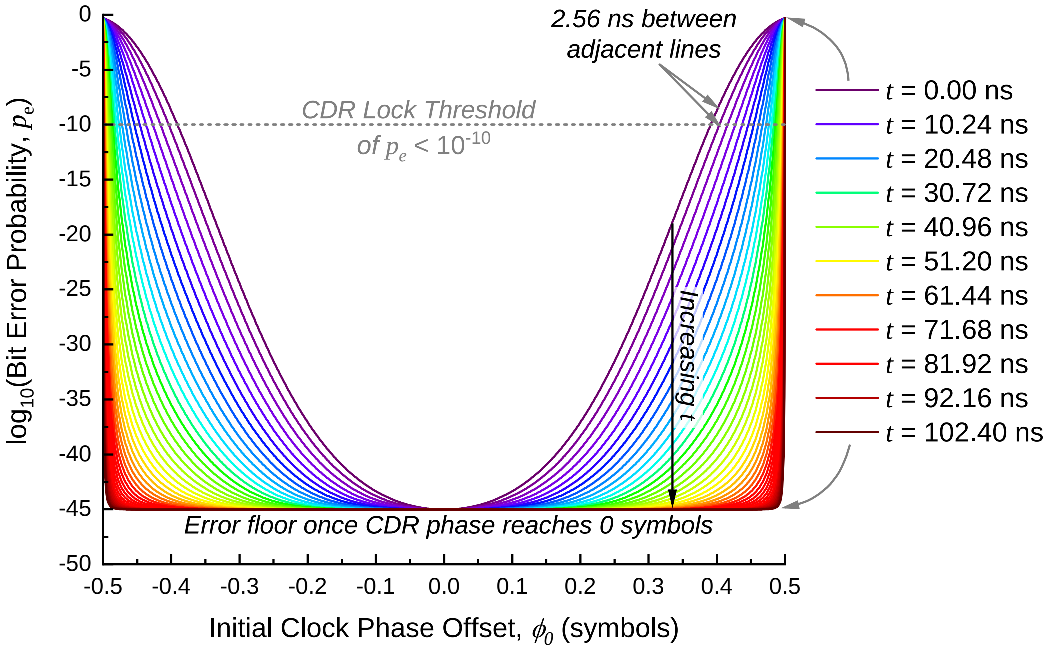

Fig. 5.6: Bit error probability as a function of initial clock phase offset. This was obtained using Equations 5.10, 5.11 and 5.12. Moving vertically from one coloured line to the nearest adjacent line represents an increase of t of 2.56 ns. The same parameters used to generate Figure 5.5 were also used to generate this figure. A CDR locking time threshold error probability of 10^{-10} is also shown.

Bit error probability is minimised when sampling occurs at the optimal sampling point for each transmitter to receiver pair, which occurs after the CDR circuit moves the sampling phase to $\phi = 0$ symbols. During the process of CDR locking, the bit error probability begins at a maximum at the beginning of packet reception at $t = 0$, before decreasing as the CDR circuit moves the clock phase sampling position towards the optimal sampling point. The initial rate-of-change of bit error probability is slow for values of ϕ_0 that are close to $|\phi| = 0.5$ due to metastability of the CDR in this clock phase offset region. At exactly $|\phi| = 0.5$, the bit error probability does not change due to CDR metastability.

A large initial clock phase offset of approximately 0.4 symbols is required to lead to initial clock phase offsets that do not result in an instantaneous CDR locking time, i.e. with an error probability that is not greater than 10^{-10} at the beginning of packet reception at $t = 0$. However, the equations obtained in this section do *not* include the impact of jitter on the sampling of the pulse shape, which will cause increases in error probability at all initial clock phase offsets. The impact of jitter will be more fully modelled in the next sub-section.

5.4.4 Evolution of Bit Error Probability within Received Data Packets (Including the Impact of Jitter on Clock Phase)

In single calibration CSA-CDR without packet clock phase tracking, as modelled in Chapter 4, no clock recovery takes place as the CDR circuit does not run except for initial clock phase offset calibration. As a consequence, the bit error probability is impacted by jitter of all frequencies. All jitter present within data arriving at the receiver is rejected, as the data is re-sampled directly using the local reference clock.

The impact of jitter is more complex in single calibration CSA-CDR with packet clock phase tracking. In this case, as shown in Figure 5.1, the CDR circuit is running continuously, only reset between packets to a constant value. At the beginning of packet reception, the CDR circuit is not locked to low-frequency jitter between the sampling clock and the data, but the CDR circuit will begin to track and then lock to this low-frequency jitter after sufficient time has passed since the beginning of packet reception. In contrast, high-frequency jitter, which occurs on timescales shorter than the measurement time interval of the CDR circuit, is rejected, in the same manner as for all jitter frequencies in single calibration synchronisation data recovery[†].

The contribution of high and low frequency jitter to the total jitter will be modelled as random independent Gaussian processes. Using the central limit theorem, the sum of the variances of the high-frequency jitter, $k^2_{\mathrm{jit-h}}$, and low-frequency jitter, $k^2_{\mathrm{jit-l}}$, may then defined as equal to the variance of the total jitter, k^2_{jit}:

$$k^2_{\mathrm{jit}} \triangleq k^2_{\mathrm{jit-h}} + k^2_{\mathrm{jit-l}} \tag{5.14}$$

For the purpose of calculations made in this chapter and later in this thesis, $k_{\mathrm{jit-l}}$ will be assumed to be 0.0210 symbols, which is measured for a synchronised Xilinx CDR in Chapter 6. Given that k_{jit} is measured to be 0.0604 symbols in Chapter 6, $k_{\mathrm{jit-h}}$ can then be calculated to be 0.0566 symbols using Equation 5.14.

[†] Jitter could also have a moderate frequency, which would cause variation of the optimal clock phase sampling position on within-packet timescales. This effect is not emulated within this analytical model due to the author's judgement that this would excessively complicate the model. This would, however, be an interesting area for further research.

If the random clock phase offset caused by high frequency jitter is defined as α, and the random clock phase offset caused by low frequency jitter is defined as β, then the probability distribution functions describing the probability density of the phase offset position as a function of phase for the high frequency jitter, $\mathrm{PDF}_{\mathrm{jit-h}}(\alpha)$ and the low frequency jitter, $\mathrm{PDF}_{\mathrm{jit-l}}(\beta)$, are then:

$$\mathrm{PDF}_{\mathrm{jit-h}}(\alpha) = \frac{1}{k_{\mathrm{jit-h}}\sqrt{2\pi}} \exp\left(\frac{\alpha^2}{2k_{\mathrm{jit-h}}}\right) \tag{5.15}$$

$$\mathrm{PDF}_{\mathrm{jit-l}}(\beta) = \frac{1}{k_{\mathrm{jit-l}}\sqrt{2\pi}} \exp\left(\frac{\beta^2}{2k_{\mathrm{jit-l}}}\right) \tag{5.16}$$

The high-frequency jitter can be modelled in a similar manner as for single calibration synchronisation data recovery without clock recovery, by performing a cross-correlation between the jitter Gaussian probability distribution function and the bit error probability arising from the pulse shape. This is performed by multiplying the probability density of each clock phase offset due to high-frequency jitter, $\mathrm{PDF}_{\mathrm{jit-h}}(\alpha)$, by the error probability with that clock phase offset due to high-frequency jitter applied, given by $p_e(\mathrm{E}[\phi(t)] + \alpha)$. $\mathrm{E}[\phi(t)]$ then gives the evolution of the expectation of the clock phase offset as a function of time since initial packet reception, where α is the additional clock phase offset caused by high-frequency jitter. This can be written as:

$$p_{e(\mathrm{post-jit-h})}(t) = \int_{-\infty}^{\infty} \mathrm{PDF}_{\mathrm{jit-h}}(\alpha) p_e(\mathrm{E}[\phi(t)] + \alpha) d\alpha \tag{5.17}$$

Figure 5.7 shows the bit error probability with high-frequency jitter accounted for as a function of time since first arrival of a packet, $p_{e(\mathrm{post-jit-h})}(t)$, and Figure 5.8 shows the bit error probability with high-frequency jitter accounted for as a function of initial clock phase offset, ϕ_0. The error floor reached when the CDR moves the sampling phase to the optimal sampling point at ϕ = 0, is increased in comparison to Figures 5.5 and 5.6. This occurs because even when sampling at the optimal sampling point, high-frequency jitter will cause some data edges to arrival with a clock phase offset from the optimal sampling point.

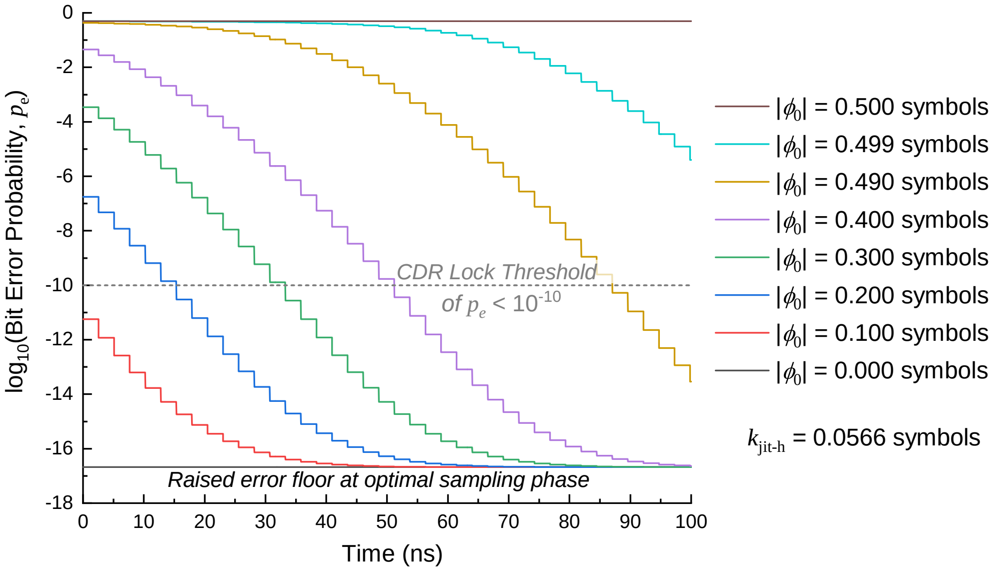

Fig. 5.7: Bit error probability with high-frequency jitter accounted for, as a function of time since the beginning of packet reception. This was obtained using Equation 5.17, using the initial clock phase offset magnitudes, $|\phi_0|$, from Figure 5.4. The error floor at the optimal sampling point is raised when compared to the no jitter case. A value of $k_{\text{jit-h}}$ of 0.0566 symbols was used.

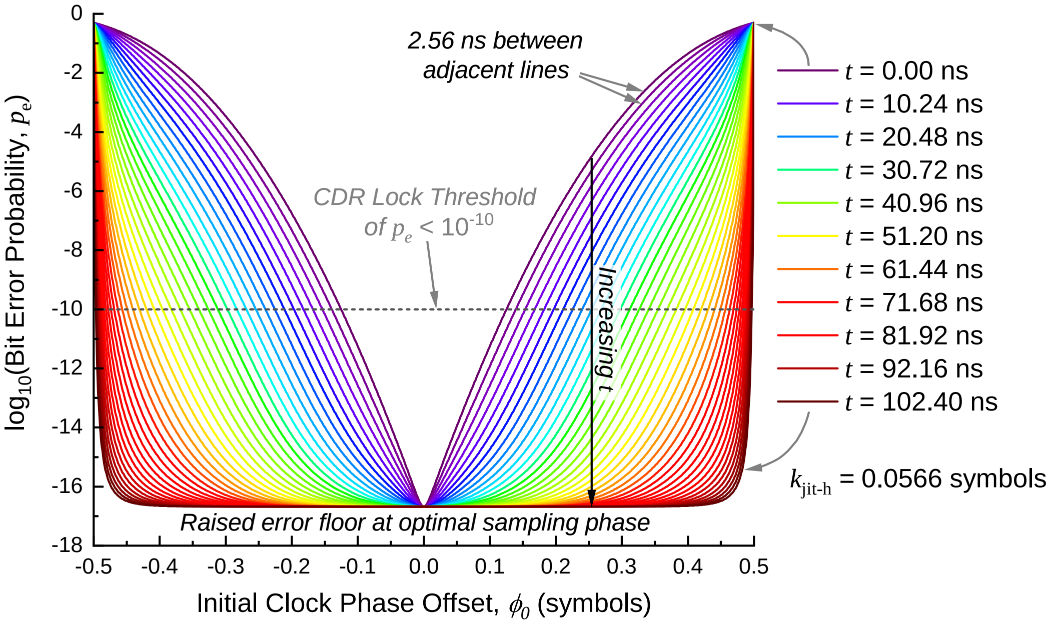

Fig. 5.8: Bit error probability with high-frequency jitter accounted for, as a function of initial clock phase offset. This was obtained using Equation 5.17. Moving vertically from one coloured line to the nearest adjacent line represents an increase of t of 2.56 ns. The error floor at the optimal sampling point is raised when compared to the no jitter case. A value of $k_{\text{jit-h}}$ of 0.0566 symbols was used.

The low-frequency jitter can be considered to be random fluctuations of the clock phase offset that occur on timescales longer than the CDR locking time. This results in initial clock phase offsets at the first packet reception that are offset by a random clock phase offset due to low frequency jitter. The impact of this on bit error probability can then be modelled by multiplying the probability density of each clock phase offset due to low-frequency jitter, $\mathrm{PDF}_{\mathrm{jit-l}}(\beta)$, by the error probability as a function of the expectation of the clock phase offset since initial packet reception, $p_e(\mathrm{E}[\phi(t,\beta))])$. An additional clock phase offset due to low-frequency jitter, β, is added to the clock phase offset at initial packet reception, ϕ_0. This can be written using the following series of equations:

$$p_{e(\mathrm{post-jit-l})}(t) = \int_{-\infty}^{\infty} \mathrm{PDF}_{\mathrm{jit-l}}(\beta) p_e(\mathrm{E}[\phi(t,\beta)])d\beta \tag{5.18}$$

where $\mathrm{E}[\phi(t,\beta))]$ is altered from Equation 5.9 by including the clock phase offset due to low-frequency jitter, β, which describes the evolution of the clock phase offset as a function of time since initial packet reception:

$$\mathrm{E}[\phi(t+dt,\beta)] \approx \begin{cases} \mathrm{E}[\phi(t,\beta)] - p\,\mathrm{erf}\left(\frac{\mathrm{E}[\phi(t,\beta)]-n}{\sqrt{2}k_{\mathrm{jit}}}\right)dt, \\ \qquad \text{if } n-\frac{1}{4} < \mathrm{E}[\phi(t,\beta)] \leq n+\frac{1}{4} \\ \mathrm{E}[\phi(t,\beta)] + p\,\mathrm{erf}\left(\frac{\mathrm{E}[\phi(t,\beta)]-n-\frac{1}{2}}{\sqrt{2}k_{\mathrm{jit}}}\right)dt, \\ \qquad \text{if } n+\frac{1}{4} < \mathrm{E}[\phi(t,\beta)] \leq n+\frac{3}{4} \end{cases} \tag{5.19}$$

where $n \in \mathbb{Z}$ and $k_{\mathrm{jit}} < \frac{1}{8\sqrt{2}}$. The initial clock phase offset at initial packet reception (at t = 0), $\mathrm{E}[\phi(0,\beta)]$, is then is given by:

$$\mathrm{E}[\phi(0,\beta)] = \phi_0 + \beta \tag{5.20}$$

Figure 5.9 shows the bit error probability with low-frequency jitter accounted for as a function of time since first arrival of a packet and Figure 5.10 shows the bit error probability with high-frequency jitter accounted for as a function of initial clock phase offset, ϕ_0. A large, Gaussian shaped error floor, which is very small in Figures 5.5 and 5.6, occurs at clock phase offsets beyond $|\phi|$ = 0.2 symbols. This error floor occurs because low-frequency jitter will cause the initial clock phase offset of some packets to be close in magnitude to, at or beyond $|\phi|$ = 0.5 symbols. This causes the CDR to lock to a neighbouring symbol period (with associated loss of frame lock) rather than the intended symbol period centred on ϕ = 0 symbols, or for the CDR phase to remain within the metastable region rather than moving to the optimal sampling point.

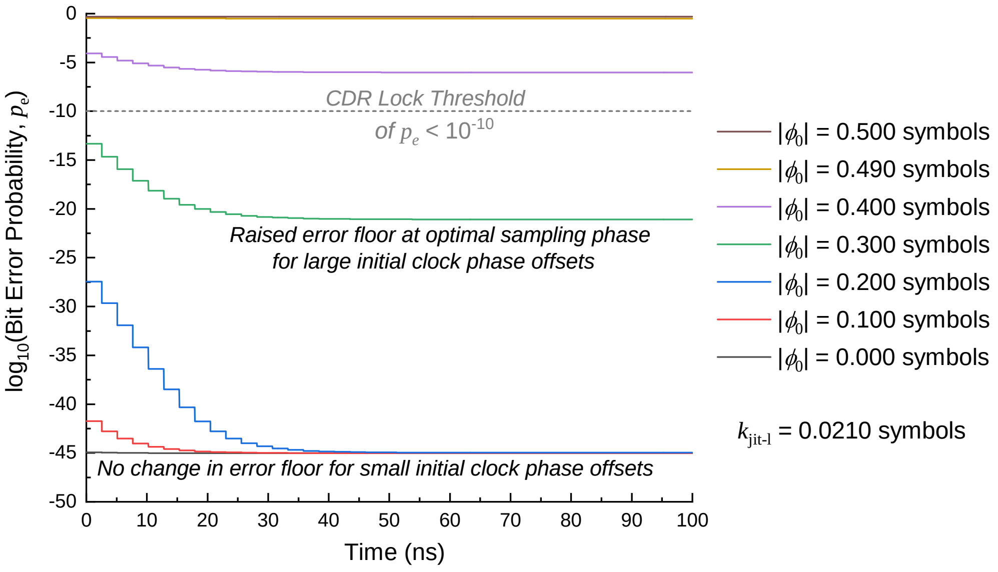

Fig. 5.9: Bit error probability with low-frequency jitter accounted for, as a function of time since the beginning of packet reception. This was obtained using Equations 5.18, 5.19 and 5.20, using the initial clock phase offset magnitudes, $|\phi_0|$, from Figure 5.4. An error floor occurs at large initial clock phase offsets, due to low frequency jitter causing the initial starting phase to sometimes fall at or beyond $|\phi|$ = 0.5 symbols. There is no change in the final error probability for small clock phase offsets. A value of $k_{\text{jit}-1}$ of 0.0210 symbols was used.

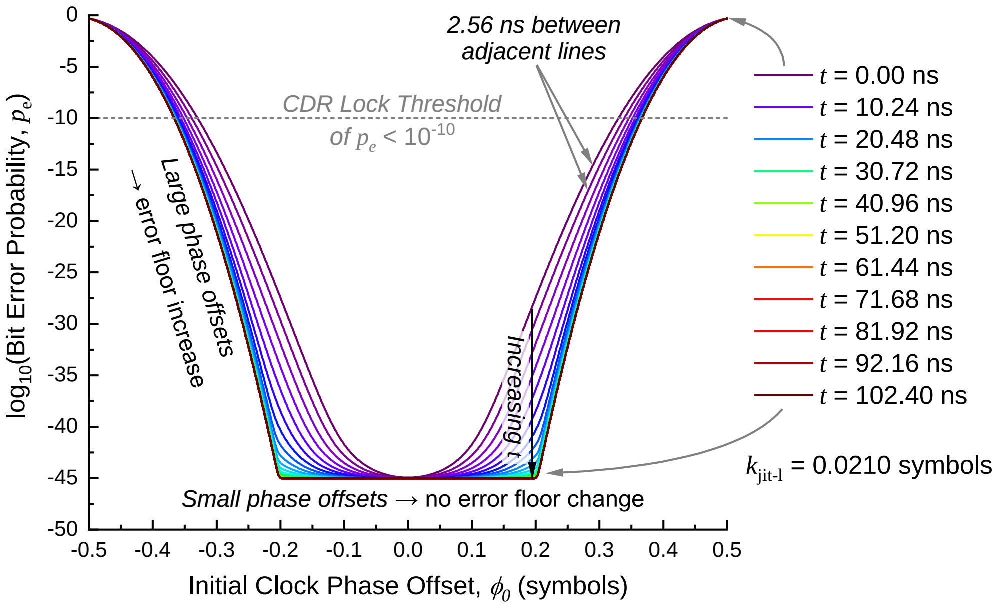

Fig. 5.10: Bit error probability with low-frequency jitter accounted for, as a function of initial clock phase offset. This was obtained using Equations 5.18, 5.19 and 5.20. Moving vertically from one coloured line to the nearest adjacent line represents an increase of t of 2.56 ns. An error floor occurs at large initial clock phase offsets, due to low frequency jitter causing the initial starting phase to sometimes fall at or beyond $|\phi|$ = 0.5 symbols. There is no change in the final error probability for small clock phase offsets. A value of $k_{\text{jit}-1}$ of 0.0210 symbols was used.

As the high-frequency and low-frequency contributions to jitter are independent random Gaussian processes, Equations 5.17 and 5.18 can be combined to give the overall bit error probability as a function of time since first packet reception, after accounting for both high-frequency and low-frequency jitter. The error probability after accounting for high and low-frequency jitter in single calibration synchronisation assisted CDR can then be modelled by the following set of equations:

$$p_{e(\text{post}-\text{jit})}(t) = \int_{-\infty}^{\infty} \text{PDF}_{\text{jit}-\text{l}}(\beta) \int_{-\infty}^{\infty} \text{PDF}_{\text{jit}-\text{h}}(\alpha) p_e(t,\alpha,\beta)\, d\alpha\, d\beta \tag{5.21}$$

where $p_e(t,\alpha,\beta)$ is given by Equation 5.10 (with t replaced with t,α,β to include the additional dependence on α and β), which describes the error probability resulting from the NRZ-OOK pulse shape including the impact of clock phase offset due to temperature change and jitter. $p_e(t,\alpha,\beta)$ is given by:

$$\begin{aligned} p_e(t,\alpha,\beta) = \frac{1}{4}\Bigg(&\operatorname{erfc}\left(\frac{I_1(t,\alpha,\beta) - RP_{\text{avg}}}{\sqrt{4qI_1(t,\alpha,\beta)f_c + 2\sigma_{ni}^2}}\right) \\ &+ \operatorname{erfc}\left(\frac{RP_{\text{avg}} - I_0(t,\alpha,\beta)}{\sqrt{4qI_0(t,\alpha,\beta)f_c + 2\sigma_{ni}^2}}\right)\Bigg) \end{aligned} \tag{5.22}$$

where q is the electronic charge of one electron, f_c is the cut-off frequency of the TIA, σ_{ni} is the input referred noise of the TIA, R is the PIN photodiode responsivity, P_{avg} is the average received optical power, and $I_1(t,\alpha,\beta)$ and $I_0(t,\alpha,\beta)$ are the positive and negative-going NRZ-OOK photocurrent pulses given in Equations 5.11 and 5.12 (with t again replaced with t,α,β to include the additional dependence on α and β). The positive and negative-going NRZ-OOK photocurrent pulses, including the additional clock frequency offset, α, due to high-frequency jitter, are then:

$$\begin{aligned} I_1(t,\alpha,\beta) = RP_{\text{avg}}\left(\frac{r_e - 1}{r_e + 1}\right)\Bigg(&\operatorname{erf}\left(\frac{\text{E}[\phi(t,\beta)] + \alpha + \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right) \\ &- \operatorname{erf}\left(\frac{\text{E}[\phi(t,\beta)] + \alpha - \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)\Bigg) + \frac{2RP_{\text{avg}}}{r_e + 1} \end{aligned} \tag{5.23}$$

$$\begin{aligned} I_0(t,\alpha,\beta) = RP_{\text{avg}}\left(\frac{r_e - 1}{r_e + 1}\right)\Bigg(2 - &\operatorname{erf}\left(\frac{\text{E}[\phi(t,\beta)] + \alpha + \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right) \\ &+ \operatorname{erf}\left(\frac{\text{E}[\phi(t,\beta)] + \alpha - \frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)\Bigg) + \frac{2RP_{\text{avg}}}{r_e + 1} \end{aligned} \tag{5.24}$$

where r_e is the transmitter extinction ratio, k_{pls} is the standard deviation in symbols of the Gaussian filter used to obtain the NRZ-OOK photocurrent pulse shape, R is the PIN photodiode responsivity and P_{avg} is the average received power.

Lastly, k_{pls} is related to the cut-off frequency of the TIA, f_c, by the symbol rate, B:

$$k_{\mathrm{pls}} = \frac{B\sqrt{\ln 2}}{2\pi f_c} \tag{5.25}$$

The expectation of the clock phase offset as a function of time since the beginning of packet reception, $\mathrm{E}[\phi(t,\beta)]$, is evaluated numerically using Equations 5.19 and 5.20.

Figure 5.11 then shows the bit error probability with high-frequency and low-frequency jitter accounted for as a function of time since first arrival of a packet, and Figure 5.12 shows the bit error probability with high-frequency jitter accounted for as a function of initial clock phase offset, ϕ_0. The two types of error floors, from high and low-frequency jitter in Figures 5.8 and 5.10 respectively, are combined in Figures 5.11 and 5.12. The combination of high and low-frequency jitter also results in a bit error probability increase from the high-frequency jitter caused error floor when the initial clock phase offset, ϕ_0 begins at the optimal sampling phase. This occurs because the initial sampling phase may be initially offset from the optimal sampling position by low-frequency jitter. The bit error probability then decreases to the high-frequency jitter caused error floor as the CDR moves the sampling phase to the optimal sampling phase at $\phi = 0$.

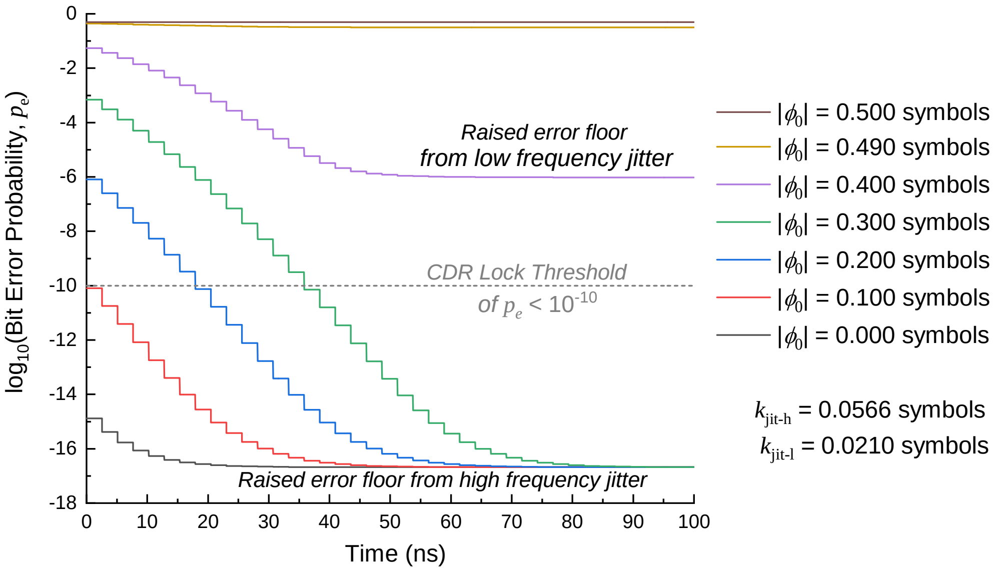

Fig. 5.11: Bit error probability with low-frequency and high-frequency jitter accounted for, as a function of time since the beginning of packet reception. This was obtained using Equation 5.18, using the initial clock phase offset magnitudes, $|\phi_0|$, from Figure 5.4. The two error floors from high and low frequency jitter combine when accounting for both types of jitter. A value of $k_{\mathrm{jit-h}}$ of 0.0566 symbols and a value of $k_{\mathrm{jit-l}}$ of 0.0210 symbols were used to generate the error probability after jitter. The total jitter, k_{jit}, resulting from the root mean square of the high and low frequency jitter contributors, was 0.0604 symbols).

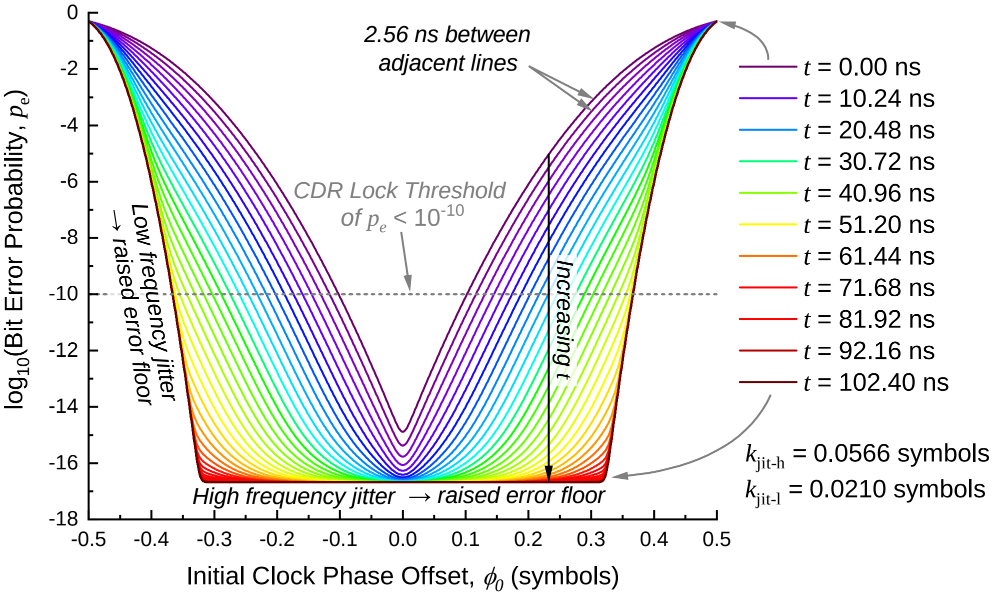

Fig. 5.12: Bit error probability with low-frequency and high-frequency jitter accounted for, as a function of initial clock phase offset. This was obtained using Equation 5.18. Moving vertically from one coloured line to the nearest adjacent line represents an increase of t of 2.56 ns. The two error floors from high and low frequency jitter combine when accounting for both types of jitter. A value of $k_{\mathrm{jit-h}}$ of 0.0566 symbols and a value of $k_{\mathrm{jit-l}}$ of 0.0210 symbols were used to generate the error probability after jitter. The total jitter, k_{jit}, resulting from the root mean square of the high and low frequency jitter contributors, was 0.0604 symbols).

5.5 Analytical Modelling of CDR Locking Time

5.5.1 Impact of Initial Clock Phase Offset

The CDR locking time as a function of initial clock phase offset can be easily obtained for single calibration CSA-CDR with packet clock phase tracking from Figure 5.12. This is achieved by evaluating the time since first packet reception, t, at which the bit error probability first falls beneath a threshold error probability, $p_{e(\mathrm{lock})}$, of 10^{-10}. Note that this is equivalent to counting the number of bit error probability curves in the previous section that are above the CDR locking time threshold in the bit error probability against initial clock phase offset figures, then subtracting one from this count (since the up-most curve is for t = 0), and then multiplying this number by 2.56 ns (which is the time, t, elapsed between each adjacent curve).

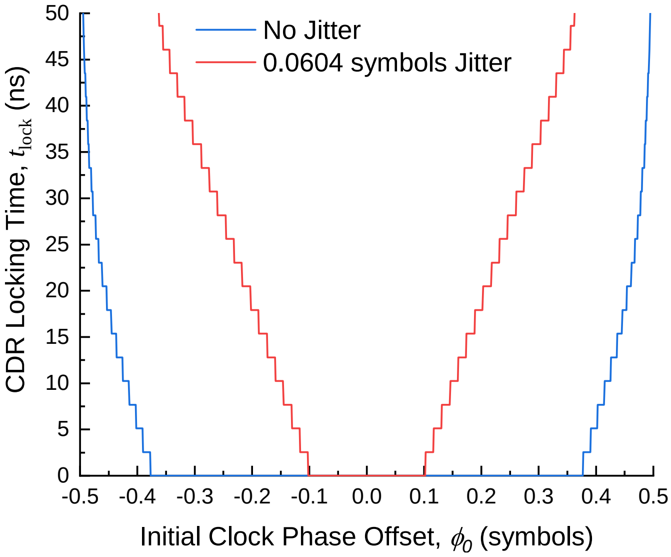

Fig. 5.13: CDR locking time against initial clock phase offset, for both an ideal case without clock jitter, and a practical case where significant clock jitter is present. These were generated from Figures 5.6 and 5.12. The range of initial clock phase offset values that result in instantaneous CDR locking time is reduced by clock jitter. A value of $k_{\mathrm{jit-h}}$ of 0.0566 symbols and a value of $k_{\mathrm{jit-l}}$ of 0.0210 symbols were used to generate the CDR locking time values with jitter included. The total jitter, k_{jit}, resulting from the root mean square of the high and low frequency jitter contributors, was 0.0604 symbols).

Figure 5.13 shows the impact of clock phase offset on CDR locking time with and without clock jitter, for an average received optical power of -10.5 dBm, the set of receiver characteristics given in Chapter 4 Section 4.4.2, and a value of $k_{\mathrm{jit-h}}$ of 0.0566 symbols and a value of $k_{\mathrm{jit-l}}$ of 0.0210 symbols (measured from a Xilinx GTY receiver CDR). For the case without jitter, CDR lock is instantaneous for initial clock phase offsets of up to $|\phi_0| < 0.377$ symbols. For the case with jitter, CDR lock is instantaneous for initial clock phase offsets of up to $|\phi_0|$ <0.100 symbols, which is greatly reduced from the without jitter case. Additionally, for the case with jitter, clock phase offsets of greater than $|\phi_0|$ = 0.35 symbols, the CDR does not lock within the packet length as the bit error probability is always greater than 10^{-10}. This occurs because low frequency jitter causes some packets to arrive with a phase offset within the receiver CDR metastable region.

5.5.2 Impact of Data Centre Temperature Change

In this subsection, the CDR locking time as a function of change in data centre temperature since clock phase calibration will be estimated for the four different data centre distance scales. Recall from Chapter 4 that the change in clock phase, $\Delta\phi$, due to the temperature change, ΔT, is given by Chapter 4 Equation 4.10 to be $\Delta\phi = 2\tau LB\Delta T$, where τ is the thermal coefficient of delay of the optical fibre used to transport the clock and data, L is the distance between nodes and B is the symbol rate.

Assuming ΔT represents the change in temperature that has occurred since the initial clock phase was calibrated to be 0 symbols, Chapter 4 Equation 4.10 can be substituted for the initial clock phase offset ϕ_0 in Equation 5.20, to give:

$$\mathrm{E}[\phi(0,\beta)] = 2\tau LB\Delta T + \beta \tag{5.26}$$

Given a fibre length L equal to the data centre distance scale, and assuming the use of loose-tube buffered SMF-28 with a thermal coefficient of delay, τ of 42.6 ps/(km·°C), Equations 5.19, 5.21, 5.22, 5.23 and 5.24 can then be used with Equation 5.26 to calculate bit error probability as a function of change in temperature, ΔT, including the impact of high and low-frequency jitter.

Figure 5.14 then shows the CDR locking time as a function of temperature change, at each of the four typical data centre length scales (Figure 5.14a, inter-rack ($\leq$7 m); Figure 5.14b, intra-cluster ($\leq$100 m); Figure 5.14c, core ($\leq$2 km) and Figure 5.14d, inter-building ($\leq$10 km)), including the impact of high and low-frequency jitter for a typical synchronised data centre FPGA transceiver. The range of acceptable temperatures for each length scale are 13.68 °C for intra-rack ($\leq$7 m), 0.957 °C for intra-cluster ($\leq$100 m), 0.0479 °C for core ($\leq$2 km) and 0.0957 °C for inter-building ($\leq$10 km).

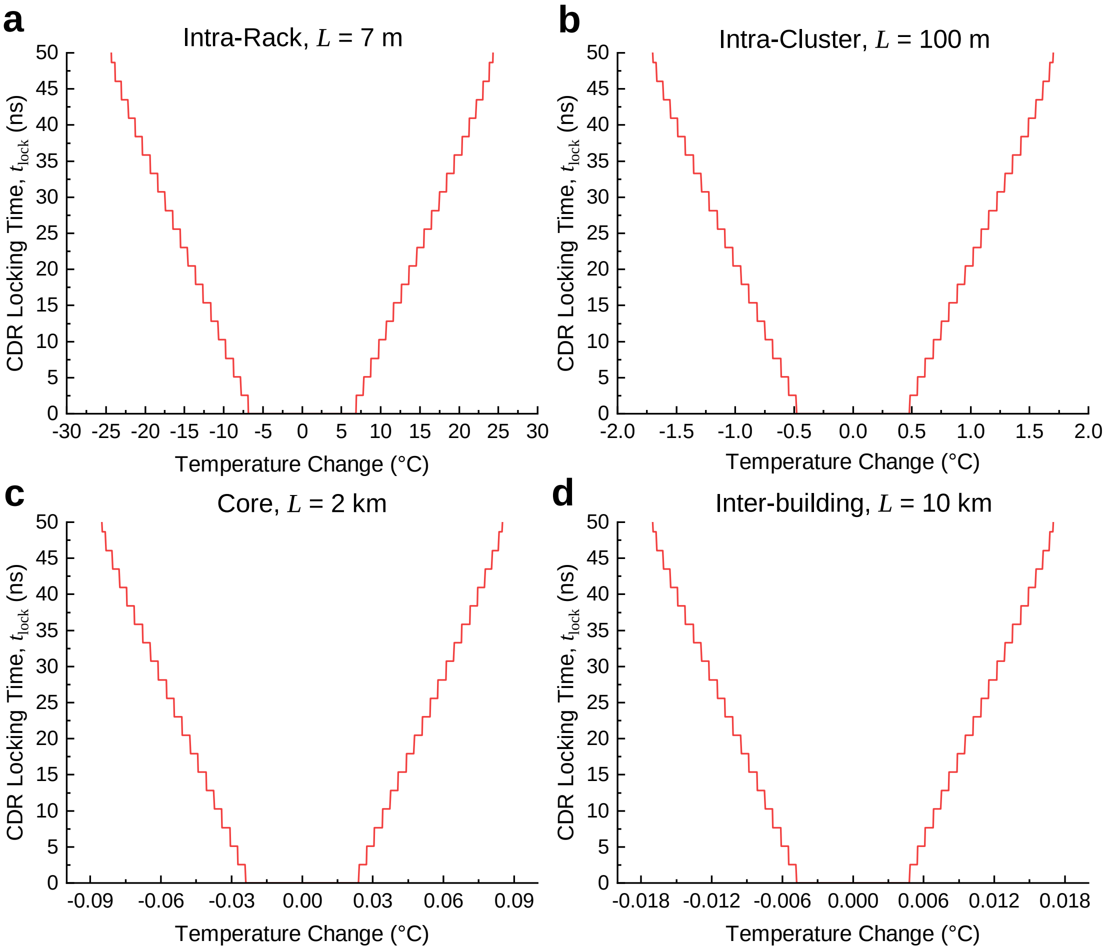

Fig. 5.14: CDR locking time against temperature change since clock phase calibration, including the impact of high and low frequency jitter. A value of $k_{\mathrm{jit-h}}$ of 0.0566 symbols and a value of $k_{\mathrm{jit-l}}$ of 0.0210 symbols were used to generate the CDR locking time values with jitter included. **a**, inter-rack (≤7 m); **b**, intra-cluster (≤100 m); **c**, core (≤2 km); **d**, inter-building (≤10 km).

5.5.3 Summary of Analytical Modelling Assumptions

A set of assumptions were made in this chapter as part of the construction of the analytical modelling, which may be summarised as:

- All the assumptions listed in Chapter 4 Subsection 4.4.3 were also made as part of the analytical modelling introduced in this chapter.

- Both the low and high frequency contributions to jitter were modelled as Gaussian. This was a reasonable assumption since both high and low frequency jitter are caused by processes that generally are Gaussian distributed. There may however also be deterministic sources of jitter, such as duty cycle distortion, that are not emulated by the analytical modelling.

- That the clock phase error output by the BB-PD is always equal to its expectation value. In practise the clock phase error output by the BB-PD will take a distribution of values. This simplification was made to minimise the complexity of the analytical modelling.

- The CDR phase is modelled as being able to take any infinitely small subdivision of one symbol. In practise, digital PI CDRs may only take a limited set of subdivisions of one symbol, for example, in steps of $\frac{1}{64}$ symbols for Xilinx UltraScale GTY CDRs. This causes a small quantisation error. This was not modelled because this effect was judged to have a small influence of bit error probability.

- The moderate frequency contribution to jitter is treated instead as either contributing to high or low frequency jitter, i.e. jitter is modelled as either occurring on timescales smaller than the phase measurement interval of the CDR (high frequency jitter), or occurring on timescales longer than the time taken to receive a packet (low frequency jitter). This decision was made to minimise the complexity of the analytical modelling.

- That the magnitude of the total jitter is small enough, $k_{\text{jit}} < \frac{1}{8\sqrt{2}}$, such that the approximation to the clock phase error of a BB-PD under the influence of Gaussian distributed jitter calculated using Equation 5.9 is within 99% of the exact analytical modelled value of clock phase error given in Appendix A Equation A.40. This is reasonable as $k_{\text{jit}} \geq \frac{1}{8\sqrt{2}}$ would cause a very large loss of signal integrity (as shown in, for example, Chapter 4 Figure 4.19d, which uses a jitter value just above this cut-off).

5.6 Discussion

As discussed in Chapter 2, worst-case data centre temperature variation of intake air for servers in a data centre rack, as given by recommendations for data centre design, may be by up to 40 °C [22]. Server exhaust temperature may vary by up to 70 °C due to local hot-spots caused by equipment such as ToR switches [68]. As shown in Figure 5.14, for the case including for the case including jitter, the CDR locking time is equal to or under 10.24 ns over a 40 °C temperature range for the within-rack case, and clock phase lock is not found over a 40 °C temperature range for all other data centre length scales. Instantaneous CDR lock is potentially possible for within-rack communication over a 27.38 °C temperature range.

For intra-rack communication, single calibration CSA-CDR with packet clock phase tracking would be potentially feasible if the optical fibre interconnecting the servers and ToR switch is laid within the temperature controlled cool aisle (i.e. on the server air intake side), and the temperature range is restricted to the A1 class of allowable temperatures (18 to 27 °C) recommended by ASHRAE [22]. In this case, the range of acceptable temperatures for these classes is within the 13.68 °C range for instantaneous CDR locking time when using single calibration CSA-CDR with packet clock phase tracking. However, as with single calibration CSA-CDR without packet clock phase tracking, this technique relies on the use of single-mode optical fibre, which would increase cost versus the short-reach copper interconnects currently used for intra-rack communication, due to the added need for optical transceivers.

Further refinement of single calibration CSA-CDR with packet clock phase tracking is required to increase its practicality. As with an approach to single calibration CSA-CDR without clock tracking, low thermal sensitivity fibre, such as HCF or MCF, could be used to reduce the clock phase shift that occurs with temperature change.

5.7 Contribution Statement

The analytical modelling presented in this chapter was conceived, constructed and numerically evaluated by K.A.C., supervised by Z.L..

Chapter 6

Clock Phase Caching for Sub-Nanosecond CDR Locking Time

6.1 Introduction

This chapter analytically models and experimentally demonstrates *clock phase caching*, an approach to CSA-CDR where the stored clock phase values are regularly updated to compensate for the changing temperature within the data centre environment, thereby limiting the worst-case clock phase offsets. This chapter begins by extending the analytical model constructed in Chapters 4 and 5 to include the impact of performing clock phase updates. Using this analytical model, the theoretical viability of the approach at intra-data centre distance scales is established. Clock phase caching is then demonstrated experimentally, achieving under 625 ps CDR locking time in a real-time FPGA-based proof-of-concept prototype. The long-term stability of the technique is evaluated, in addition to the tolerance of the technique to rate-of-change of temperature, clock jitter and small optical power. Finally, on the basis of the results of this evaluation, the scalability of the technique to supporting large-port count optical switches is explored. The analytical modelling presented in this chapter is presented for the first time. The experimental work presented in this chapter was first published as a post-deadline conference paper at ECOC 2018 [19] and was then published as a journal paper in Nature Electronics [1].

6.2 Clock Phase Caching Concept

Clock phase caching extends the single calibration CSA-CDR approach explored in Chapters 4 and 5. A clock phase cache is located within each transceiver, containing a set of values (one value per receiver) corresponding to the phase shifts that need to be applied to the synchronised clock before each packet is sent from its transmitter. At start-up every transmitter exchanges a packet with all receivers connected through the optical switch in the data centre. As in single calibration CSA-CDR, every receiver then measures the phase offset of the received packet and feeds back this information to the transmitter to populate its phase cache. However, unlike in single calibration CSA-CDR, in clock phase caching this process is then repeated periodically to compensate for clock phase drifts that occur due to temperature variation causing fibre time-of-flight change. A single clock phase update between a transmitter and receiver is shown in Figure 6.1.

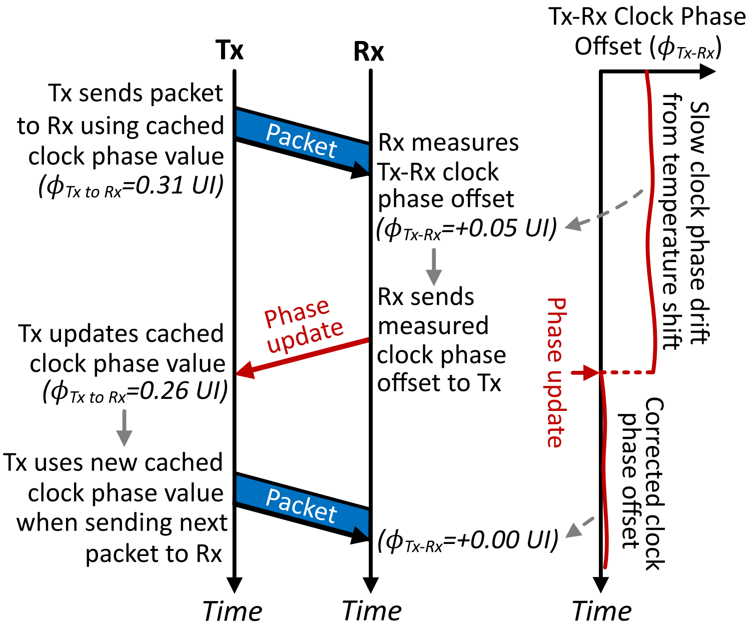
Tx
Rx
Tx-Rx Clock Phase Offset ($\phi_{Tx\text{-}Rx}$)
Tx sends packet to Rx using cached clock phase value ($\phi_{Tx\ to\ Rx}$=0.31 UI)
Packet
Rx measures Tx-Rx clock phase offset ($\phi_{Tx\text{-}Rx}$=+0.05 UI)
Slow clock phase drift from temperature shift
Rx sends measured clock phase offset to Tx
Phase update
Phase update
Tx updates cached clock phase value ($\phi_{Tx\ to\ Rx}$=0.26 UI)
Tx uses new cached clock phase value when sending next packet to Rx
Packet
($\phi_{Tx\text{-}Rx}$=+0.00 UI)
Corrected clock phase offset
Time
Time
Time

Fig. 6.1: Operational principle of clock phase caching, showing one clock phase update for one transmitter-receiver pair. In clock phase caching, clock phase updates are performed at a slow rate across all transmitter-receiver pairs. This aligns the clock phase of all data packets arriving at each receiver irrespective of origin, simplifying sub-nanosecond clock and data recovery. UI, unit interval, equal to one symbol period; Tx, transmitter; Rx, receiver. Figure adapted from [1].

6.2.1 Important Considerations for Implementing Clock Phase Caching

The extension of single calibration CSA-CDR to include regular clock phase updates potentially introduces transmission overhead on the optical switch. This transmission overhead originates from the time required to be used by receivers to perform the clock phase offset measurements for all transmitters connected to each receiver. In contrast, single calibration CSA-CDR does not introduce transmission overhead as the phase calibration process only occurs at startup; it does not repeat. This transmission overhead must be balanced against the accuracy of measurement of the clock phase offset, in the presence of jitter present between the sampling clock and the clock embedded within the data of incoming packets. A longer clock phase offset measurement time should act to average over sampling jitter. A balance must therefore be achieved that ensures that the clock phase offset measurement and update process is both *1)* sufficiently fast, and *2)* sufficiently accurate.

6.3 Analytical Modelling of Clock Phase Caching

In this section, bit error probability at a receiver in a clock phase cached optical switch will be modelled by extending the analytical model established in Chapters 4 and 5.

6.3.1 Clock Phase Cached Bit Error Probability

Consider that the expectation of the clock phase offset at the start of incoming packets, E$[\phi(0,\beta)]$, in the presence of a temperature caused clock phase offset, $2\tau LB\Delta T$, and a low frequency jitter caused clock phase offset, β, is given by Chapter 5 Equation 5.26:

$$\text{E}[\phi(0,\beta)] = 2\tau LB\Delta T + \beta \tag{6.1}$$

where τ is the transmission fibre thermal coefficient of delay, L is the distance between servers/switches connected to the optical switch, B is the symbol rate and ΔT is the temperature shift since the last calibration of the clock phase offset.

If t^* is defined as the time elapsed since the last calibration of the clock phase offset, $\Delta T(t^*)$ can then be further defined as the change in fibre temperature after t^*. If the change in ΔT is assumed to be negligible on packet timescales, t, which are on the order of up to 100 ns, then the expectation of the clock phase offset at the start of incoming packets as a function of time elapsed since calibration, E$[\phi(0,t^*,\beta)]$, can be calculated from:

$$\text{E}[\phi(0,t^*,\beta)] = 2\tau LB\Delta T(t^*) + \beta \tag{6.2}$$

In a CSA-CDR approach using clock phase caching, the critical environmental factor is not the absolute temperature change in the data centre, but the magnitude of the change since the last clock phase update. The temperature change since the last clock phase update, $\Delta T(t^*)$, can be calculated using the rate of change of temperature, $\frac{dT}{dt}$, through the following integral:

$$\Delta T(t^*) = \int_0^{t^*} \frac{dT}{dt} dt \tag{6.3}$$

If the time, t^*, since the last clock phase update is assumed to be short, such as on timescales of milliseconds to up to 10 seconds, then considering the significant bulk of air mass within the data centre environment, the rate of change of temperature could therefore reasonably be assumed to be constant on these timescales. Under this assumption, Equation 6.3 simplifies to:

$$\Delta T(t^*) \approx t^* \frac{dT}{dt} \tag{6.4}$$

Equation 6.4 can then be substituted into Equation 6.2 to give the expectation of the clock phase offset at the start of incoming packets as a function of time elapsed since the last clock phase update, $\mathrm{E}[\phi(0, t^*, \beta)]$:

$$\mathrm{E}[\phi(0, t^*, \beta)] \approx 2\tau LBt^* \frac{dT}{dt} + \beta \tag{6.5}$$

The clock phase update rate, f_ϕ, can then be defined as the rate at which clock phase offset measurement and updates are performed for all Tx-Rx pairs connected to the optical switch. At the instant of a clock phase update, t^* = 0. At the time instant immediately prior to the next clock phase update, t^* = $1/f_\phi$. The minimum and maximum initial clock phase offsets, ϕ_{min} and ϕ_{max} respectively, are then therefore:

$$\phi_{\mathrm{min}} = 0 \tag{6.6}$$

$$\phi_{\mathrm{max}} = \frac{2\tau LB}{f_\phi} \frac{dT}{dt} \tag{6.7}$$

Recall that $p_{e(\mathrm{post-jit})}(t)$ from Chapter 5 is the bit error probability as a function of time since the beginning of packet reception, including the impact of clock jitter. If the impact of the linear increasing initial clock phase offset from the data centre temperature change is included, $p_{e(\mathrm{post-jit})}(t)$ then becomes $p_{e(\mathrm{post-jit})}(t, t^*)$, with the expectation of the clock phase offset at the beginning of packet reception given by Equation 6.5. The overall phase cached bit error probability, as a function of time since the beginning of packet reception, t, can then be obtained by calculating the mean bit error probability, $p_{e(\mathrm{ph-cached})}(t)$, between the moment in time that the clock phase offset updates occur, at $t^* = 0$ and the time instant immediately prior to the next clock phase offset update $t^* = 1/f_\phi$, which can be performed by integration:

$$p_{e(\mathrm{ph-cached})}(t) = f_\phi \int_0^{\frac{1}{f_\phi}} p_{e(\mathrm{post-jit})}(t, t^*)dt^* \tag{6.8}$$

Equivalently, the mean bit error probability at each time since the beginning of packet reception, t, can also be calculated by integrating the bit error probability, $p_{e(\mathrm{post-jit})}(t, \phi_0)$, in terms of initial clock phase offset, ϕ_0, between the minimum, ϕ_{min}, and maximum, ϕ_{max}, initial clock phase offset values of $\phi_0 = 0$ and $\phi_0 = 2\tau LB\frac{dT}{dt}/f_\phi$ respectively. The mean bit error probability is then:

$$p_{e(\mathrm{ph-cached})}(t) = \frac{f_\phi}{2\tau LB}\frac{dt}{dT}\int_0^{\frac{2\tau LB}{f_\phi}\frac{dT}{dt}} p_{e(\mathrm{post-jit})}(t, \phi_0)d\phi_0 \tag{6.9}$$

An example integral between these two initial clock phase offsets is illustrated in Figure 6.2, which uses the bit error probability curves from Chapter 5 Figure 5.12, which show the bit error probability as a function of initial clock phase offset and time since the beginning of packet reception, including the impact of high and low frequency jitter. Each individual bit error probability against initial clock phase offset curve is integrated over to find the phase cached bit error probability as a function of time since the beginning of packet reception. Figure 6.3 and Figure 6.4 then show the result of these integrals. The bit error probability is reduced at each maximum phase cached initial clock phase offset, ϕ_{max}, versus the same valued initial clock phase offsets in Figure 6.2. This reduction in bit error probability arises from the effect of averaging larger bit error probabilities arising from large initial clock phase offsets immediately prior to a clock phase update, with smaller bit error probabilities arising from small initial clock phase offsets immediately after a clock phase update.

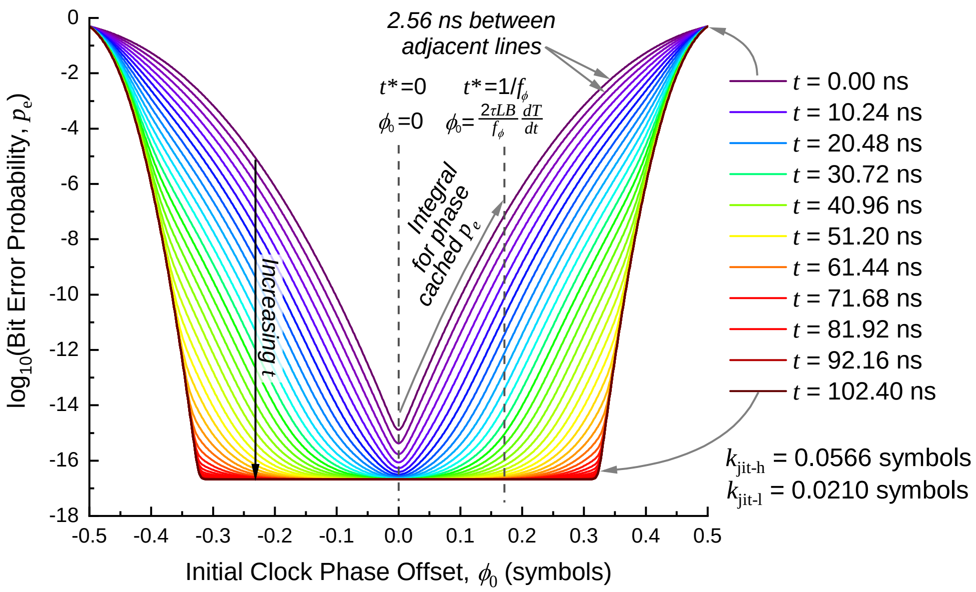

Fig. 6.2: Illustration of the integrals used to evaluate bit error probability in a clock phase cached receiver. If the rate of change of temperature is assumed to be linear, then the overall clock phase cached bit error probability can be calculated by finding the mean bit error probability between the minimum initial clock phase offset immediately following a clock phase update and the maximum initial clock phase offset immediately prior to the next clock phase update. In this example, the bit error probability curves to be integrated over are from Chapter 5 Figure 5.12.

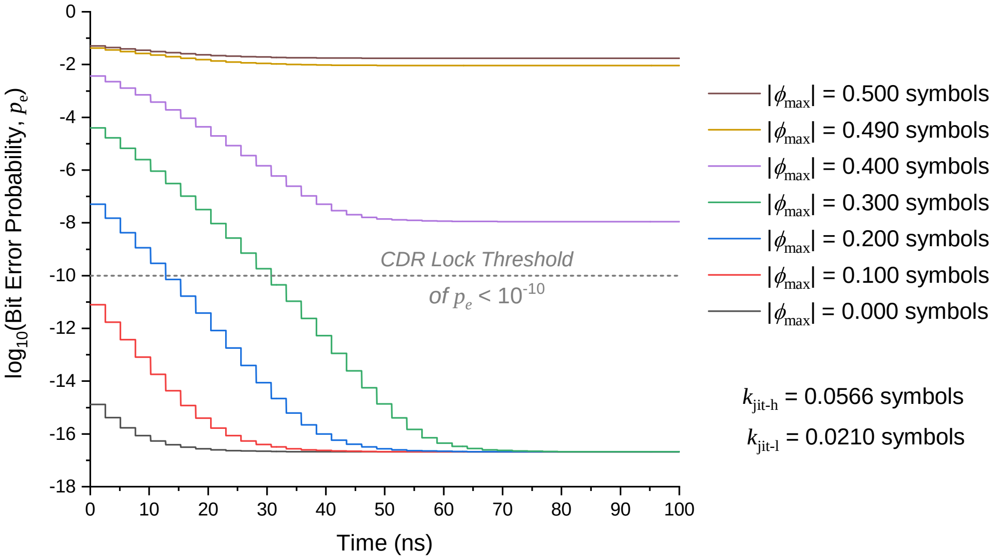

Fig. 6.3: Bit error probability in a clock phase cached receiver, as a function of time since the beginning of packet reception. A CDR locking time threshold of 10^{-10} is shown, after which the CDR is considered to have locked.

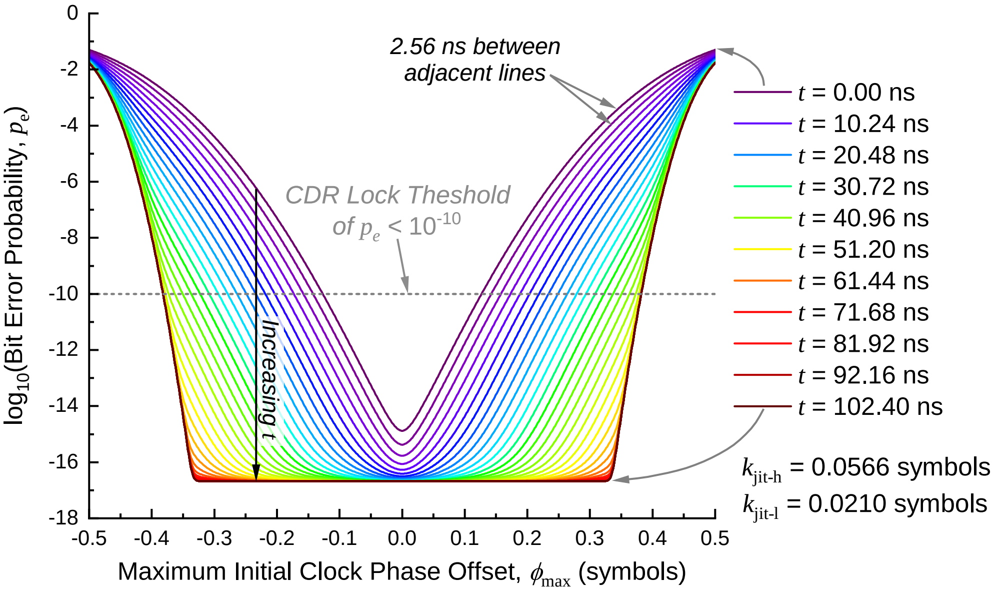

Fig. 6.4: Bit error probability in a clock phase cached receiver, as a function of maximum initial clock phase offset. The bit error probability is reduced at each maximum phase cached initial clock phase offset, $\phi_{\max}$, versus the same valued initial clock phase offsets in Figure 6.2. A CDR locking time threshold of 10^{-10} is shown, after which the CDR is considered to have locked.

6.3.2 CDR Locking Time in a Clock Phase Cached Receiver

The CDR locking time, t_{lock}, as a function of maximum initial clock phase offset, ϕ_{max}, can then be calculated from Figure 6.4 by determining the time since the beginning of packet reception, t, at which the bit error probability falls to beneath $p_{e(\text{lock})}$. Figure 6.5 shows the CDR locking time as a function of maximum initial clock phase offset, ϕ_{max}, with a threshold bit error probability, $p_{e(\text{lock})}$, of 10^{-10}. In this case, with the post-jitter bit error probability generated using realistic receiver performance characteristics detailed in Chapters 4 and 5, if the maximum initial clock phase offset of arriving packets can be constrained to be under $\approx \pm 0.136$ symbols, then the CDR locks instantaneously as the overall clock phase cached bit error probability at the beginning of packet reception is already under the threshold of 10^{-10}.

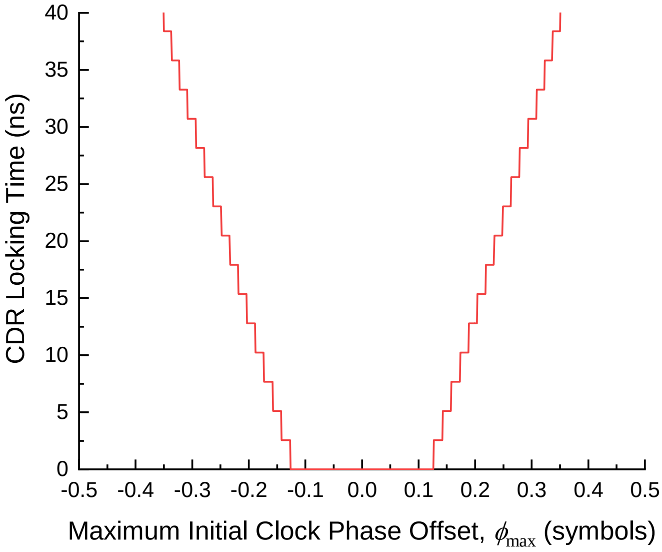

Fig. 6.5: Clock and data recovery locking time in a clock phase cached receiver, as a function of maximum initial clock phase offset. A CDR locking time threshold of 10^{-10} was used to generate the CDR locking time curve. If the maximum initial clock phase offset immediately prior to clock phase updates can be constrained to be under $\approx \pm 0.136$ symbols, then the CDR achieves locks instantaneously as the bit error probability at the beginning of packet reception is already under the threshold of 10^{-10}.

6.3.3 Effect of Rate-of-Change of Temperature on CDR Locking Time

The CDR locking time, t_{lock}, as a function of rate of temperature, $\frac{dT}{dt}$, and clock phase update rate, f_{ϕ}, in a clock phase cached receiver, can then be calculated using the values for CDR locking time against maximum clock phase offset, ϕ_{max}, shown in Figure 6.5. The calculation is performed by using the values from Figure 6.5 in combination with Equation 6.7, assuming a distance between nodes, L, of 2 km, a fibre thermal sensitivity, τ, for loose-tube buffered SMF-28 of 42.6 ps/(km·°C) [94] and a symbol rate, B, of 25 GBaud. These parameters match those that are used in the experimental evaluation of clock phase caching performed later in this chapter.

Figure 6.6 then shows the result of this calculation. The CDR locking time was evaluated for three clock phase update rates, 10 Hz, 6.25 Hz and 2.5 Hz. An increase in clock phase update rate causes a proportional increase in the rate-of-change of temperature at which the CDR locking time rises to above 0 ns. All three of these clock phase update rates are sufficiently fast to maintain 0 ns CDR locking time under worst-case rates of change of data centre temperature (0.0556 °C/s [22]).

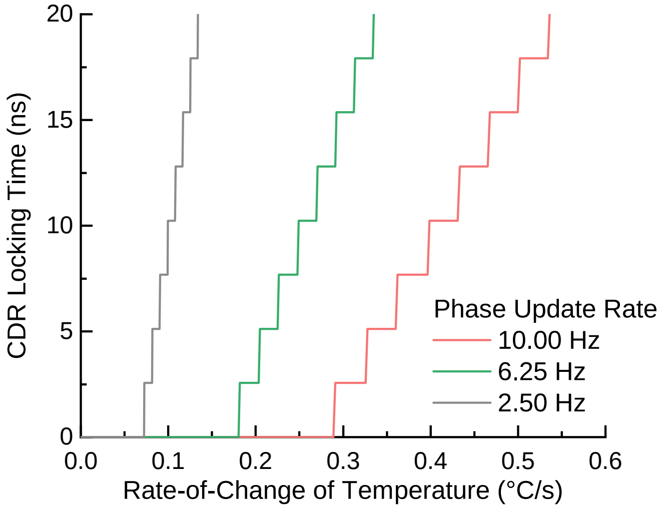

Fig. 6.6: CDR locking time as a function of rate of temperature and clock phase updates in a clock phase cached receiver. The CDR locking time as a function of rate of change of temperature was evaluated for three clock phase update rates, 10 Hz (red line), 6.25 Hz (green line) and 2.5 Hz (grey line). A CDR locking time threshold of 10^{-10} was used to generate the CDR locking time curves. A distance between nodes of 2 km, a fibre thermal sensitivity for loose-tube buffered SMF-28 of 42.6 ps/(km·°C) [94] and a symbol rate of 25 GBaud were assumed.

6.3.4 Effect of Sampling Clock Gaussian Jitter on CDR Locking Time

The impact of sampling clock jitter on CDR locking time can also be quantified. To perform this evaluation, the fibre temperature was assumed to be constant, i.e. no clock phase shift due to temperature occurs, such that $\phi_{\max}$ is always 0 symbols, and so any degradation in error probability purely arises from an increase in jitter. The minimum jitter frequency is assumed to be significantly greater than a slow clock phase cache update rate of 1 to 10 Hz, such that clock phase caching is unable to compensate for the jitter. Chapter 5 Equations 5.21 to 5.25 were then used, assuming a constant high frequency contribution to jitter, $k_{\textbf{jit-h}}$, of 0.0566 symbols, while the low frequency contribution to jitter, $k_{\textbf{jit-l}}$ was varied from 0 to 0.075 symbols. All other parameters used to generate Figure 6.5 are again used in this calculation. The total jitter and CDR locking time was calculated using the above equations. This method was chosen to emulate a similar experimental result that is presented later in this chapter.

Figure 6.7 shows the outcome of this calculation. The CDR locking time remains instantaneous up to a total random jitter, k_{jit}, of 0.076 symbols, at which point the CDR locking time rapidly worsens as the total random jitter is increased. This degradation arises from clock jitter causing some packets to have initial clock phase offsets to fall within regions of the pulse shape that have a bit error probability that is worse than 10^{-10}. The CDR then takes time to shift the clock phase back into a region with a bit error probability that is equal to or under 10^{-10}. This result demonstrates the need to minimise jitter in optically-switched systems that use CSA-CDR as an approach to perform burst-mode CDR, such that the deviation in the initial clock phase offset caused by low-frequency jitter is minimised.

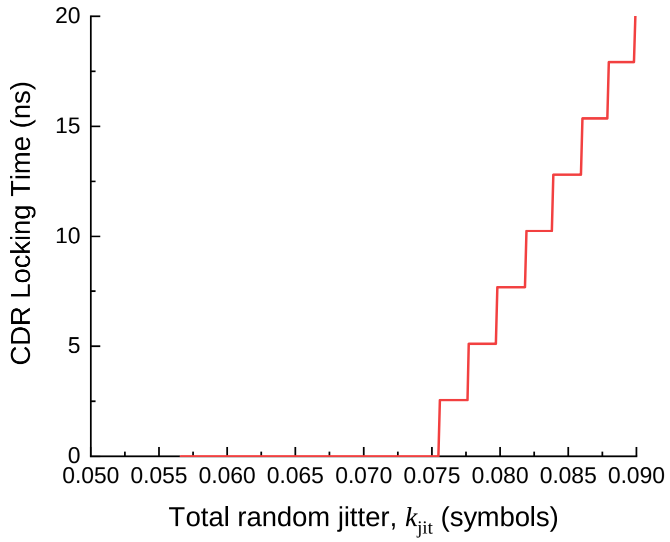

Fig. 6.7: Clock and data recovery locking time in a clock phase cached receiver, as a function of total jitter. A CDR locking time threshold of 10^{-10} was used to generate the CDR locking time curve. Increased jitter causes degradation of CDR locking time.

6.3.5 Minimum Required Rate of Clock Phase Updates

If the threshold maximum initial clock phase offset, $\phi_{\max\,(\text{thresh})}$, is defined as the largest value of $\phi_{\max}$ at which the CDR locking time remains instantaneous (i.e. the bit error probability is 10^{-10} or under from the beginning of packet reception), then Equation 6.7 can be used to calculate the minimum required clock phase update rate, $f_{\phi(\min)}$, to compensate for a given rate of change of temperature.

$$f_{\phi(\min)} = \frac{2\tau LB}{\phi_{\max(\text{thresh})}} \frac{dT}{dt} \tag{6.10}$$

It is convenient to define the proportionality constant, k, between rate of change of temperature and the minimum required rate of clock phase updates to maintain instantaneous CDR lock for a given Tx-Rx pair:

$$k \triangleq \frac{2\tau LB}{\phi_{\max(\text{thresh})}} \tag{6.11}$$

$$f_{\phi(\min)} = k\frac{dT}{dt} \tag{6.12}$$

Equation 6.10 can be used with the value of $\phi_{\max(\text{thresh})}$ of 0.126 symbols and typical data centre transmission parameters to estimate the required clock phase update rate for clock phase caching operating on a Tx-Rx link connected through an optical switch. Firstly, a distance between nodes, L, of core data centre transmission scales, of 2 km, is assumed. A fibre thermal coefficient of delay, τ, of 42.6 ps/(km·°C), matching the measured thermal coefficient of delay of loose-tube buffered SMF-28 [94], is assumed. Lastly, in Chapter 2, the recommended worst-case rate of change of data centre temperature, $\frac{dT}{dt}$, given by ASHRAE, was 20 °C per hour, or 0.0556 °C/s [22].

Given these parameters, the theoretical value of k can be calculated to be 33.8 $°\text{C}^{-1}$ using Equation 6.11. The minimum required rate of clock phase updates to compensate for a 0.0556 °C/s rate of change of temperature is then therefore 1.74 Hz, calculated using Equation 6.12. The estimated required rate of clock phase updates to cope with worst-case rates of change of temperature within a data centre environment is therefore still very slow.

6.3.6 Throughput Overhead

Clock phase caching must be scalable to support large-port count data-centre optical switches. During the process of performing a clock phase update, a Tx-Rx pair is forced to communicate, even when it might not have data to transmit between the pair. If an approach to scheduling is used where end-points connected to the optical switch request grants for packet transmission [62, 63], this acts to introduce throughput overhead on the throughput of the optical switch. The time spent performing clock phase updates each second, and therefore the throughput overhead from clock phase caching, o, can be calculated as follows:

$$o = \sum_{n=1}^{N} f_{\phi(n)}(t_{\text{meas}} + t_{\text{update}}) \tag{6.13}$$

where N is the total number of nodes, $f_{\phi(n)}$ is the clock phase update rate for each node pair, t_{meas} is the time taken to measure the clock phase offset between each transmitter to receiver pair and t_{update} is the length of time each transmitter spends transmitting a packet to carry each phase update value.

The phase update rate for each link, $f_{\phi(n)}$, is in turn given by Equation 6.11. Substituting Equation 6.11 into Equation 6.13 then gives:

$$o = \sum_{n=1}^{N} \frac{2\tau_n L_n B_n}{\phi_{\max(\text{thresh},n)}} \left(\frac{dT}{dt}\right)_n (t_{\text{meas}} + t_{\text{update}}) \tag{6.14}$$

where τ_n, L_n, B_n, $\phi_{\max(\text{thresh},n)}$, $(\frac{dT}{dt})_n$ are the same quantities as described for Equation 6.10 but for each Tx-Rx link, n, connected through the optical switch.

The worst-case overhead, $o_{\max}$, then occurs when all of the above quantities take worst-case, identical (for each n) values. The sum in Equation 6.14 then reduces to:

$$o_{\max} = N f_{\phi}(t_{\text{meas}} + t_{\text{update}}) = \frac{2N\tau LB}{\phi_{\max(\text{thresh})}} \frac{dT}{dt}(t_{\text{meas}} + t_{\text{update}}) \tag{6.15}$$

6.3.7 Theoretical Viability

The very slow estimated required worst-case clock phase update rate, f_ϕ, of 1.74 Hz, to cope with recommended worst-case rates-of-change of temperature within the data centre across a 2 km distance, is crucial for the viability of clock phase caching. Consider a worst-case scenario consisting of the following conditions: all Tx-Rx pairs are connected by 2 km of data fibre and 2 km of data fibre; all Tx-Rx pairs experience a worst-case rate-of-change of fibre temperature of 0.0556 °C/s simultaneously; and the optical switch has 10,000 nodes. Under these conditions, all Tx-Rx pairs must update their phase values at a rate of 1.74 Hz to maintain an instantaneous CDR locking time. If $t_{\text{meas}} + t_{\text{update}}$ can restricted to approximately 2 μs, the overall analytically calculated worst-case overhead would be approximately 3.5%.

A short clock phase measurement time of approximately 2 μs would not be achievable using discrete components and devices, but it might be achievable using an FPGA receiver based approach. In Xilinx UltraScale GTY transceiver CDRs, it is possible to output the the CDR phase offset value at a high rate of 400 MHz. This could potentially be used to enable a 2 μs clock phase measurement and update time. Clock phase caching could therefore be practical with an FPGA-based approach.

6.3.8 Summary of Analytical Modelling Assumptions

The assumptions made in the analytical modelling shown in this chapter may be summarised as:

- All the assumptions listed in Chapter 4 Subsection 4.4.3 and Chapter 5 Subsection 5.5.3 were also made as part of the analytical modelling introduced in this chapter.

- That the fibre temperature changes linearly, which should be true on clock phase update timescales of 0.1 to 1 second timescales (corresponding to a clock phase update rate of 10 to 1 Hz) due to the large bulk of air within the data centre.

- That the measurement of the clock phase offset is perfectly accurate. In practise there will be some inaccuracy of this measurement, which is minimised by taking a large number of samples of the clock phase offset.

- That the same worst-case clock phase update rate is required for all links, assuming worst-case values for fibre length and rate-of-change of temperature. In practise a smaller clock phase update rate could be used for links that are of shorter length than the worst-case, or experience a smaller rate of change of temperature than the worst-case. This would lead to a smaller overall optical switch throughput overhead.

6.4 Proof-of-Concept Experimental Setup

Having determined through analytical modelling that clock phase caching is feasible, a proof-of-concept experimental system was then built, shown in Figure 6.8, which used three clock synchronised and phase cached FPGAs. This was used to investigate the performance of clock phase caching in a 2 × 1 optical switching network.

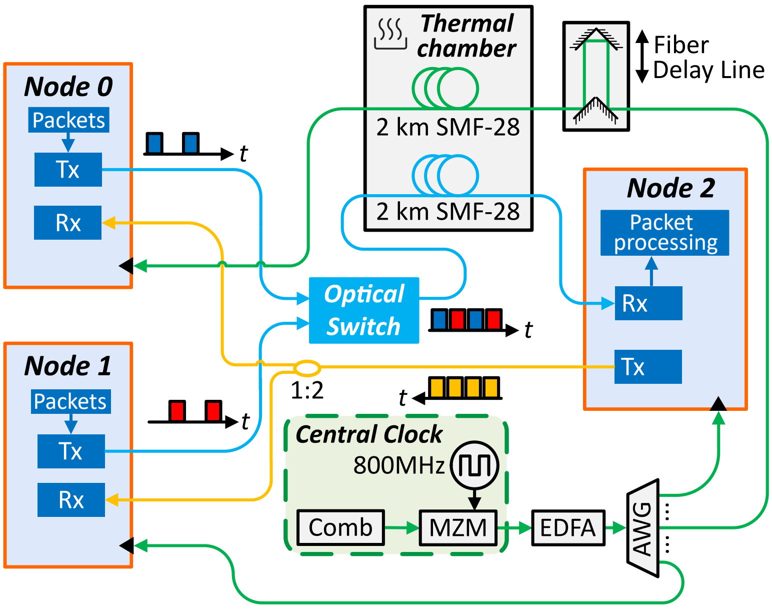

Fig. 6.8: Proof-of-concept experimental demonstration of clock phase caching operating on a 2-to-1 optical switch. 2 km clock fibre and 2 km data fibre was placed in a thermal chamber allowing the impact of temperature on clock phase caching to be investigated. The signals in these fibres counter-propagate to study the effect of worst-case rates-of-change of clock phase due to temperature on clock phase caching. Tx, transmitter; Rx, receiver; EML, externally modulated laser; MZM, Mach-Zehnder modulator; PD, photodiode; EDFA, erbium-doped fibre amplifier; AWG, arrayed-waveguide grating. Figure adapted from [1].

6.4.1 Data Packet Transmission and Optical Switch Implementation

As shown in Figure 6.8, Node 0 and Node 1 transmit 128-byte OOK modulated packet payloads embedded in 60 ns packets at 25.6 Gb/s, via externally modulated lasers (EMLs), which consist of 1550 nm carriers modulated with data using 35 GHz bandwidth electro-absorption modulators (EAMs). The packets from Node 0 and Node 1 propagate to Node 2 through a 2×2 $LiNbO_3$ Mach-Zehnder modulator (MZM) optical switch with a 200 ps switching time. After the switch, alternate packets from Node 0 and Node 1 propagate through 2 km of SMF-28 and are attenuated to -10.5 dBm before reaching Node 2 to emulate power budgets of intra-data-centre transmission standards [10]. The 2×2 optical switch was driven by a 130 ns period (two 60 ns packets plus two 5 ns inter-packet gaps) square wave clock signal from a Xilinx UltraScale GTY transceiver (18 GHz bandwidth) from FPGA Node 1.

Since the experimental focus of this work is on CDR, the following experimental simplifications were made to minimise unnecessary experimental complexity: A 2×2 optical switch was used because the clock phase shift changes occur in the optical fibre rather than the optical switching elements, and therefore the results demonstrated in the rest of this chapter would remain applicable to larger scale optical switches. Both transmitters are configured to output packets continuously with no central or edge scheduling, to avoid AC-coupling signal degradation, which would be avoidable in a commercial system with DC-coupled receivers. Time synchronisation of packets from the two transmitters is performed manually, which in practise would be performed using a time synchronisation algorithm such as White Rabbit [103].

6.4.2 Establishment of Optical Clock Frequency Synchronisation

Establishing clock frequency synchronisation of all nodes is crucial for implementing CSA-CDR, which then simplifies the phase recovery step of CDR locking. One approach to achieving frequency synchronisation would be to distribute a clock modulated onto multiple optical wavelengths, which are then split passively with star couplers and distributed to reach all nodes connected to an optical switch. An example of this approach was demonstrated in this experiment.

Frequency synchronisation of the three FPGA nodes was achieved by modulating an 800 MHz reference clock onto a 24 tone optical frequency comb with 25 GHz spacing with an MZM. The optical reference clock was distributed to all three nodes for frequency synchronisation, emulating distributed frequency synchronisation techniques such as Sync-E [83]. The 1st, 12th and 24th tones of the frequency comb were distributed to Nodes 0, 1 and 2 respectively. The frequency comb used was an optoelectronic comb source consisting of a 27 dBm continuous wave (CW) light source emitting at 1555 nm, a phase modulator and an MZM [104]. Both the phase modulator and the MZM were driven by a 25 GHz radio frequency source, generating 25 GHz spacing comb 2 dB power flatness. The generated comb was modulated and subsequently amplified to 18 dBm by an Erbium-doped fibre amplifier (EDFA), yielding an average power of about 3 dBm per tone. These optical clock signals were filtered by an AWG and detected by 18 GHz photodiodes with TIAs. The lowest receiver power required for each modulated tone was -11 dBm. The use of a 1:8 optical splitter was emulated by attenuating the optical power from -0.5 dBm (power per tone after AWG) to -11 dBm, indicating that the system used in this experiment can optically synchronise 192 nodes from a single clock source.

In practice, relatively low-speed photodiodes (e.g. 5 GHz bandwidth, which were unavailable at the time of the experiment) might provide higher sensitivities because of their slightly higher responsivity (e.g. about 1 A/W). The number of reachable nodes can be easily increased to 3072 if a high power EDFA is used to amplify the modulated clock to 30 dBm. The clock signal between the AWG and Node 1 additionally travels through 2 km of SMF-28. Clock distribution does not necessitate the use of a frequency comb: conventional laser arrays with 50 GHz or 100 GHz channel spacing could also be used. Nevertheless, a frequency comb provides a compact source with well-defined wavelength spacing which may ease thermal and wavelength management in a clock synchronised network with potentially lower operating expenses. In addition, wideband and high optical signal-to-noise ratio (OSNR) frequency combs such as parametric combs [105] and thin-film $LiNbO_3$ [106] can potentially allow the system to scale to more than 10,000 nodes.

6.5 Clock Phase Caching Experimental Implementation

The clock phase cached transceiver architecture shown in Chapter 3 Figure 3.8 was implemented in the three FPGAs. To implement clock phase caching, as shown in Figure 6.1, Node 2 periodically measured the clock phase offset using the FPGA CDR module and subsequently sent the clock phase offset values back to Node 0 and Node 1 via the link shown in yellow, which emulated duplex interconnection. Further details of the FPGA implementation of clock phase caching are given in Appendix B. Two photographs of the experimental setup are shown in Appendix C.

6.5.1 Measurement of Transmitter to Receiver Clock Phase Offset

For the overhead of clock phase caching to be small, the clock phase offset measurement process must complete on microsecond timescales. This was achieved using the digital PI CDR within the FPGA receivers. For each clock phase update, transmitter to receiver clock phase offset was measured by averaging the FPGA receiver raw CDR phase across the 2$^{\text{nd}}$ and 3$^{\text{rd}}$ packet sequences, and then averaging the clock phase offset of 32 packets separated by 1 μs, taking a total of 2.08 μs per update (including inter-packet gaps). The receiver CDR phase interpolator spans 128 evenly-spaced phase steps split across 2 symbols (resolution of 64 phase steps per symbol). The raw CDR phase was output at a rate of 400 MHz, and this phase was equal to the clock phase shift applied to the receiver reference clock to keep it aligned with the bit edges in incoming data, which was achieved within the CDR by using a phase interpolator to measure the ratio between the number of times the clock edge occurs before and after the data edge in 64 bit intervals, and shifting the receiver reference clock such that this ratio equals 50%. A single packet was sent back to the transmitter node along the return link to update the clock phase offset at the transmitter once the clock phase offset had been calculated at the receiver. No oversampling was used. The built-in Xilinx UltraScale GTY transceiver receiver-side continuous time linear equaliser (CTLE) filter was used.

6.5.2 Implementation of Transmitter Clock Phase Shift

Implementing clock phase shift at the transmitter is essential for performing transmitter side CSA-CDR. This is because different paths through an optical switch to different receivers require different phase shifts to ensure that packets arrive at all receivers with the same clock phase. These differences in required phase shift could arise from different fibre lengths and fibre temperatures. As illustrated in Figure 6.9, the transmitter must shift clock phase between the transmission of each outgoing packet to compensate for these clock phase differences. This clock phase shift must complete during the guard band between outgoing data packets, which includes the time required to reconfigure the optical switch. This necessitates that the transmitter clock phase shift must complete on timescales equal to or under the time taken to reconfigure the optical switch (ideally sub-nanosecond). This process minimises the CDR clock phase shifts shown in Chapter 5 Figure 5.1.

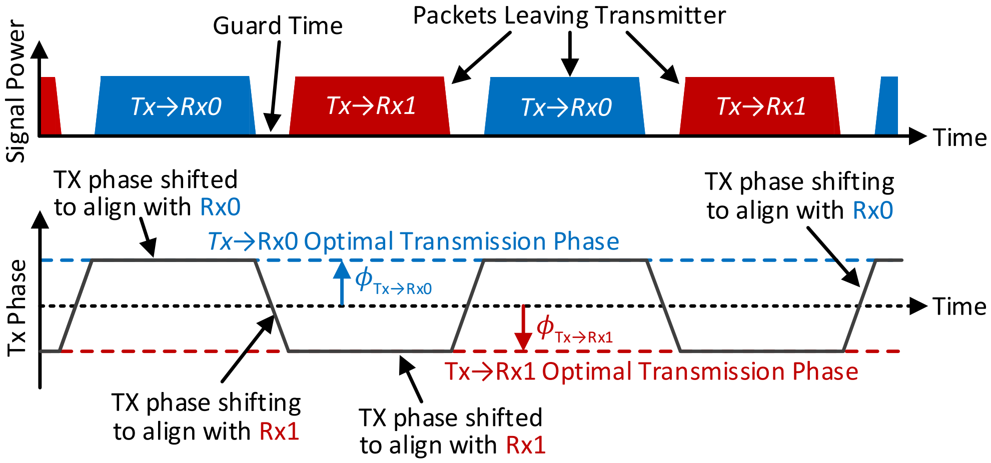

Fig. 6.9: Transmitter clock phase shift at a transmitter, Tx, to align packets with the different reference clocks of two receivers, Rx0 and Rx1. The transmitter clock phase must be shifted during the guard band (which includes the time for optical reconfiguration) between each outgoing packet to ensure the correct alignment with each receiver reference clock. Each transmitter to receiver pair is associated with its own unique clock phase offset, in this case $\phi_{\mathrm{Tx}\rightarrow\mathrm{Rx0}}$ and $\phi_{\mathrm{Tx}\rightarrow\mathrm{Rx1}}$. These unique clock phase offsets change with phase shift due to fibre temperature change and jitter.

The transmitter phase shift was implemented using the built-in Xilinx UltraScale GTY transceiver phase interpolator parts per million (PIPPM) clock phase interpolator. It consisted of a standard clock phase rotator that took a half rate clock from the output of the FPGA transceiver LC-Tank quad PLL, and output one clock phase selected from 128 evenly spaced phase steps split across 2 symbols (resolution of 64 steps per symbol). The selected clock phase was then used to drive the serial clock of the parallel in serial out (PISO) converter in the FPGA transmitter. The phase interpolator phase output was controlled by a PIPPM control circuit built into the transceiver.

The PIPPM control circuit allowed the transmitter clock phase to be shifted by up to $\frac{15}{64}$ symbols every transmitter parallel clock cycle, which was 400 MHz in this proof-of-concept experiment. Within the 5 ns interpacket gap, immediately prior to packet transmission to a receiver, the phase interpolator selected phase was changed to a clock phase value equivalent to the cached clock phase value for communication with the receiver. Shifting by half a symbol required two FPGA parallel clock cycles which took 5 ns total, hence leading to a required 5 ns gap, followed by a further $\frac{2}{64}$ symbol shift at the beginning of the next packet. The $\frac{2}{64}$ symbol shift, if required, could alternatively have been performed at the end of the previous packet. Shifting by greater than half a symbol was achieved by shifting clock phase in the reverse direction by up to half a symbol. Using this method, the transmitter can shift to any required clock phase.

The interpacket gap resulting from the transmitter phase shift could be reduced to sub-nanosecond by directly using the clock phase cached value to change the selected phase interpolator phase, bypassing the PIPPM control circuit in the FPGA transceiver. This would remove the phase shift delay associated with the PIPPM control circuit, thereby enabling the phase of the transmitter clock to switch to any arbitrary phase in sub-nanosecond time, which would fall within the 1 ns required interpacket gap required to allow for optical switching time.

Additionally, instead of performing the clock phase shifts at the transmitter as performed in this experiment, the receiver could instead measure, cache and apply the correct clock phase shifts at the receiver. This alternative approach has the advantage of simplicity and slightly reduced transmission overhead (no return path to the transmitter is required to update the clock phase values, removing the small t_{update} contribution to transmission overhead), but carries the disadvantage of requiring receivers to be aware of packet origin in advance (i.e. all transmitters and receivers must receive the optical switching schedule, which would double the optical switch control plane bandwidth required for transmitting the schedule to nodes connected to an optical switch). Transmitter clock phase shifting was implemented in this experiment because of this limitation, and because shifting the receiver clock phase on nanosecond timescales was not possible in Xilinx UltraScale GTY transceivers (but would be possible in an ASIC implementation of CSA-CDR).

6.5.3 Data Packet Structure and Measurement of CDR Locking Time

An experimental packet structure was used to measure CDR locking time and transmit the phase information that needs to be sent from the receiver to transmitter in clock phase caching, shown in Figure 6.10, contains three De Bruijn sequences [107], each of 2^9 bits length. Each sequence contained a PRBS-9 sequence of length 511 bits, generated using a standard polynomial of $x^9 + x^5 + 1$ [108], with an additional 0 added following the end of the sequence. A media access control (MAC) layer protocol (46 bits) was embedded in the third sequence for frame alignment, clock phase offset feedback and packet identification. The full details of the implementation of the MAC layer protocol are given in Appendix B. In a full system demonstration of clock phase caching, clock phase measurement packets (such as those used in this experimental demonstration) would be scheduled and transmitted periodically between standard Ethernet packets. In the prototype implementation, for the purposes of simplifying implementation complexity (since the aim of this experimentation was to demonstrate the proof-of-principle of clock phase caching), all packets transmitted were allowed to be potentially used for both clock phase measurement and data transmission, avoiding the need to implement a packet scheduler. The second and third sequences in the packets were used to measure clock phase offset when a clock phase update was required.

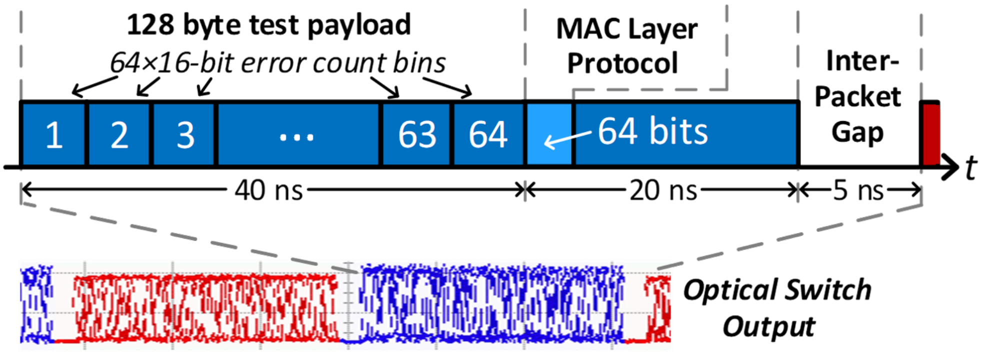

Fig. 6.10: The structure of the packets used in the clock phase caching proof-of-concept experiment. The packets consist of three 512 bit de Bruijn sequences, with the first 64 bits of the third sequence replaced with a MAC layer protocol used for indicating packet source and destination, and communicating measured clock phase offsets. The first and second 512 bit sequences are used for measuring data recovery time and bit errors. The second and third 512 bit sequences are used for measuring clock phase offset. This is performed by recording the phase output of the FPGA receiver CDR circuit in 64 bit intervals. Figure adapted from [1].

To assess the performance of phase caching, the 1st and second sequences acted as the 128 byte payload and were divided into 64×16 bit bins. The number of bit errors falling in each bin across all packets arriving at the receiver was recorded in real-time over 1 s intervals. As shown in Figure 6.11, CDR locking time was calculated as the 1st bin in the packet with a BER of under 10^{-10}, if all following bins also had a BER of under 10^{-10}. If, for example, the BERs of the first three bins are higher than 10^{-10}, and the fourth and subsequent had a BER less than 10^{-10}, the CDR locking time would therefore be 2.5 ns. If all bins had a BER of under 10^{-10}, then the CDR locking time was determined to be less than 625 ps. BER was calculated by summing errors collected across all 64 bins. A threshold BER of 10^{-10} was chosen as this enabled one BER data point to be acquired in a reasonable timespan of approximately 7 minutes, while remaining at a BER threshold at which a low-overhead FEC could be used.

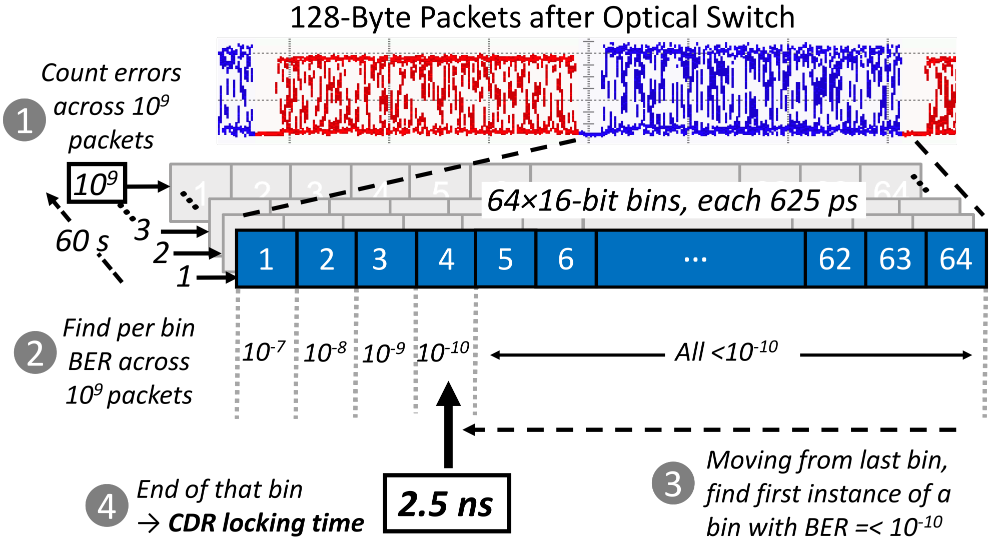
128-Byte Packets after Optical Switch
1 Count errors across 10^9 packets
10^9
3
2
1
60 s
64×16-bit bins, each 625 ps
1 2 3 4 5 6 ··· 62 63 64
2 Find per bin BER across 10^9 packets
10^{-7} 10^{-8} 10^{-9} 10^{-10}
All <10^{-10}
3 Moving from last bin, find first instance of a bin with BER =< 10^{-10}
4 End of that bin → CDR locking time
2.5 ns

Fig. 6.11: The process used to measure CDR locking time in the clock phase caching proof-of-concept experiment. Errors falling into each of the 64 16 bit bins are counted across 10^9 received packets. Per bin BER is then calculated. The CDR locking time was then calculated as the first bin in the packet with a BER of under 10^{-10}, if all following bins also have BER of under 10^{-10}. BER was calculated by summing errors collected across all 64 bins. Figure adapted from [1].

6.6 Emulation of Data Centre Environmental Conditions

The tolerance of clock phase caching to challenging environmental conditions that may occur in a practical optically-switched data centre environment needed to be investigated to assess the practicality of the technique. The clock phase drift resulting from change of temperature must be kept within tolerable boundaries to maintain sub-nanosecond CDR locking time. The magnitude of this worst-case drift is theoretically directly proportional to the optical fibre length, the rate-of-change of temperature, and is theoretically inversely proportional to the rate at which the clock phase cache values are updated. Decreased optical power and increased clock jitter, which would both act to reduce the acceptable range of clock phase offsets, also would theoretically degrade the performance of clock phase caching. This section details how the effect of these environmental conditions was experimentally evaluated.

6.6.1 Optical Fibre Length

2 km of SMF-28 for the data path and 2 km of SMF-28 for the clock path was placed in a thermally controlled chamber. These lengths are equal to the longest fibre lengths standardised for within data centre building usage [10]. When switched on, the temperature began at 25 °C and increased approximately linearly between 30 and 50 °C at a rate of 0.11 °C/s. The system was configured such that the signals propagate in opposite directions in the data and clock fibres. Variation in temperature therefore caused the phase shift in the two fibres to be additive, resulting in an overall phase shift at the receiver equal to that experienced using 4 km SMF-28 (about 160 ps/°C), allowing the worse-case rate of clock phase shift in synchronous intra-data-centre interconnection to be investigated. This contrasts with the case where clock and data signals propagate in the same direction, where balanced changes in both fibres due to temperature variation would cause the phase shifts to cancel at the receiver.

6.6.2 Rate-of-Change of Temperature

To evaluate the clock phase update rate required to maintain a CDR locking time of under 625 ps (under 16 symbols) for a given rate of change of temperature across $2{\times}2$ km single mode optical fibre, a 166 ps, voltage-controlled free space optical delay line was introduced between the AWG and the clock input of Node 1. The fibre delay line was driven with a saw-tooth waveform to experimentally emulate the rate of change of phase that occurs for a given rate of change of temperature in SMF-28. The saw-tooth waveform frequency was swept through a series of values equal to that theoretically experienced by a loose tube single mode fibre of 4 km length at a series of different rate of change of temperatures.

This rate-of-change of phase applied to the clock path was equal to $\frac{dT}{dt} \cdot \tau \cdot L$, where $\frac{dT}{dt}$ is the rate-of-change of temperature, τ is the thermal coefficient of delay of loose-tube buffered SMF-28 of 42.6 ps/(km·°C) [94] and L is the worst-case clock path length within a data centre environment of 2×2 km. 10^{11} bits per 16 bit bin were collected for each rate-of-change of temperature. To avoid the need to manually re-perform time synchronisation of the two transmitters, the switch was biased into the crossbar state such that packets only from Node 1 arrive at the receiver. To prevent the receiver CDR remembering the clock phase of incoming data between successive packets, the receiver CDR phase was reset within the 5 ns inter-packet gap.

6.6.3 Clock Jitter Generation and Measurement

To investigate the impact of jitter on clock phase caching, jitter was applied to the 800 MHz electronic clock source shown in Figure 6.8 by frequency modulating the clock source with a 1 MHz sinusoidal and white noise voltage waveforms of different amplitudes. The experimental setup used to do this is shown in Figure 6.12. The peak-to-peak sinusoidal jitter applied to the clock was calculated after the clock output and before the MZM using an radio frequency spectrum analyser by measuring the ratio of the power of the 800 MHz clock tone to the power of its ±1 MHz side-tones. The random jitter amplitude of the clock after frequency modulation with white Gaussian noise jitter was measured using a digital communications analyser. 10^{11} bits per 16 bit bin were collected for each applied jitter amplitude. To prevent the receiver CDR remembering the clock phase of incoming data between successive packets, the receiver CDR phase was reset within the 5 ns inter-packet gap.

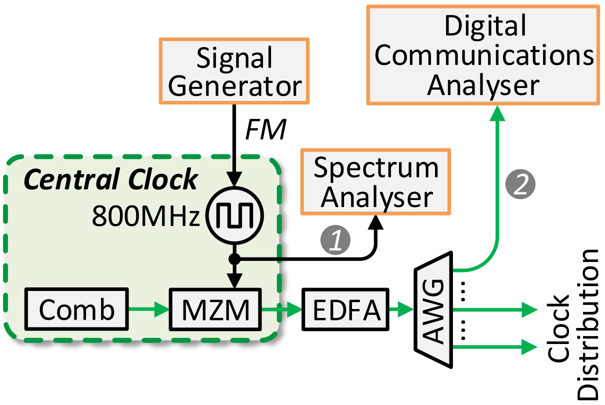

Fig. 6.12: Clock jitter generation and measurement. A signal generator applied jitter to the 800 MHz clock source by driving the frequency modulation port of the clock source. Components used for jitter injection and measurement are outlined in orange. ***1***, measurement of the peak-to-peak injected sinusoidal jitter. ***2***, measurement of the random jitter amplitude of the clock after frequency modulation with white Gaussian noise jitter. FM, frequency modulation; MZM, Mach-Zehnder modulator; EDFA, erbium-doped fibre amplifier; AWG, arrayed-waveguide grating..

6.7 Experimental Results

6.7.1 Tolerance to Clock Phase Offset

To investigate the impact of clock phase offset on CDR locking time, under steady state temperature, the 2×2 optical switch in Figure 6.8 was biased to transmit packets only from Node 0 to Node 2. The transmitter output clock phase was then moved in steps of $\frac{1}{64}$ symbols, and the CDR locking time was measured for each transmitter clock phase. 10^{11} bits per 16 bit bin were collected for each CDR locking time measurement. The average received optical power for all measurements was -10.5 dBm. Figure 6.13 shows the outcome of this investigation. There was a range of 0.20 symbols over which the CDR locking time is equal or under 625 ps (under 16 symbols), and no errors were detected. This under 625 ps CDR locking time range matches the instantaneous CDR locking time range of 0.20 symbols shown in the analytically modelled equivalent plot in Chapter 5 Figure 5.13. The CDR locking time outside of this region in Figure 6.13 increased linearly at approximately the same rate as that shown in Chapter 5 Figure 5.13, and the BER degraded accordingly.

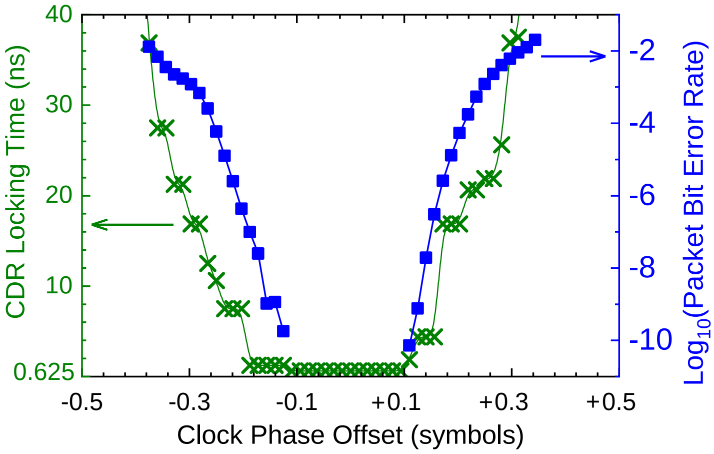

Fig. 6.13: Impact of clock phase offset on CDR locking time (green crosses) and BER (blue squares) under steady state temperature. The transmitter clock phase was moved in steps of $\frac{1}{64}$ across a range of one symbol. B-spline fits are shown as guides to the eye. Figure adapted from [19].

6.7.2 Long-Term Stability

To demonstrate the reliability of the technique, a measurement of CDR locking time was performed over a 48 hour period in a laboratory environment where the temperature fluctuated within 5 °C, as shown in Figure 6.14. During this long-term stability test, cooling failure was emulated by turning off the laboratory air conditioner (Figure 6.14a), which led to a 2 °C temperature rise at 17 hours followed by a slow room temperature increase during the 17th and 23rd hours. After that time point, the air conditioner was switched on, which cooled the room back to 18.5 °C. The receiver continuously monitored the clock phase shift of the two transmitters, as shown in Figure 6.14b. The CDR locking time was determined by finding the first error free 16 bit bin by averaging 9.2×10^8 packets over 60 s. When clock phase caching was enabled, the system was error-free over the whole 48 hours, resulting in a CDR locking time of under under 625 ps (under 16 symbols), as shown in Figure 6.14c. When phase caching was disabled a quick degradation in BER occurred, due to temperature drift, resulting in an increase in CDR locking time to over 40 ns within 4 minutes.

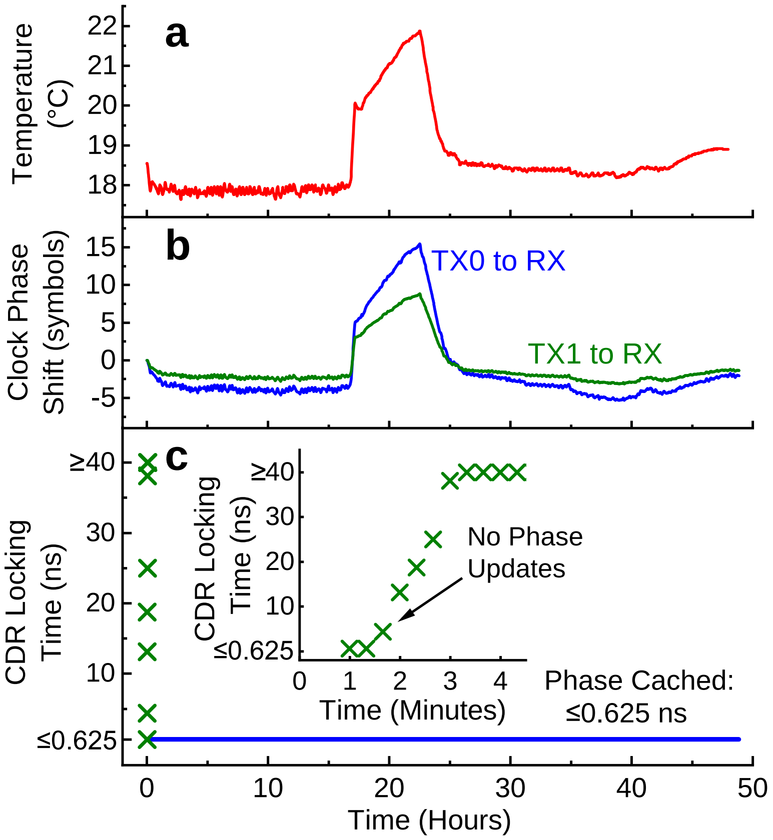

Fig. 6.14: Stability of clock phase caching over 48 hours of measurement: a, Recorded ambient temperature, with the air conditioner switched off between off between the 17th hours and 23rd hours; **b**, Recorded receiver clock phase shift for packets originating from each transmitter; **c**, Receiver CDR locking time. Blue line, with clock phase caching; Green crosses, no clock phase caching. Figure adapted from [1].

6.7.3 Resilience to Rapid Temperature Change

The magnitude of the worst-case rate-of-change of temperature within a production data centre was evaluated by monitoring the inlet and exhaust temperature for a rack in a production data centre with evaporative cooling over 228 days in 5-minute intervals. The largest temperate variation observed was 9 °C; assuming the worst-case, that this temperature variation occurs across adjacent 5 minute measurement intervals, it translates to a rate of change of 0.03 °C/s. This is also consistent with recommendations for industrial data centre design [22].

Figure 6.15 shows the performance of clock phase caching under a rapid 0.11 °C/s rate-of-change of temperature, induced on $2{\times}2$ km of SMF-28 using the thermal chamber. Figure 6.15a and Figure 6.15b show the measured temperature in the temperature chamber and the resulting total phase shifts. The 0.11 °C/s increase of temperature, which is approximately 3 times greater than was observed in a production data centre, resulted in a 16 ps per second change in fibre time-of-flight. Figure 6.15c shows the measured CDR locking time with phase caching enabled at a rate of 10 Hz. Even with the rapid change of temperature, no errors were observed, and the CDR locking time was under 625 ps (under 16 symbols). When clock phase caching was initially ran, and then switched off during the 0.11 °C/s rate of temperature increase, a loss of clock phase alignment occurred in less than one second due to the temperature induced change in fibre time-of-flight.

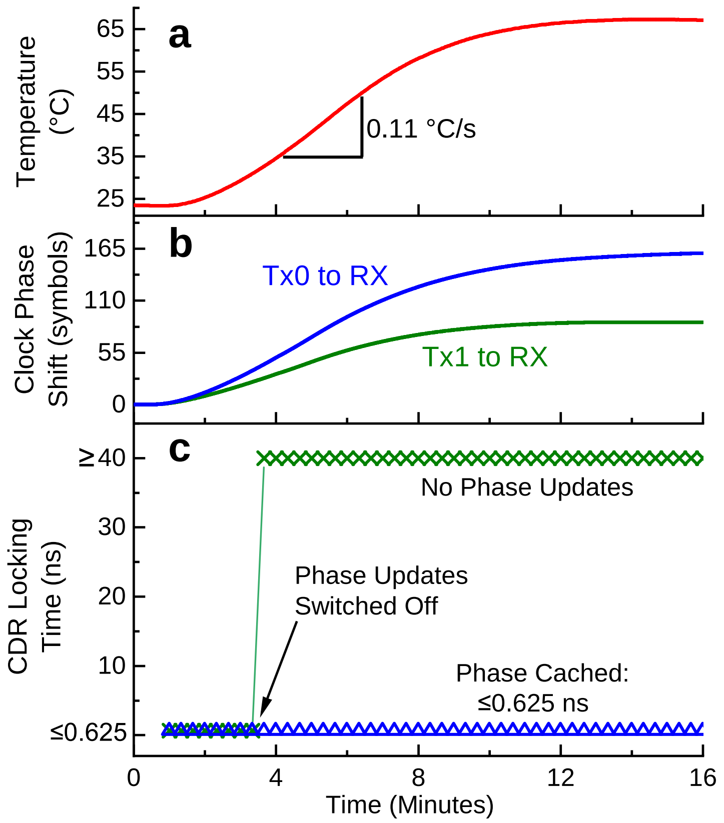

Fig. 6.15: Stability of clock phase caching under a rapid 0.11 °C/s rate of temperature change. This rate is over 3× greater than the worst-case rate of change of temperature that was observed in a production cloud data centre: **a**, Temperature within thermally controlled chamber; **b**, Recorded clock phase shift for packets originating from each transmitter; **c**, Receiver CDR locking time. Blue triangles, with clock phase caching; Green crosses, clock phase caching turned off during the rapid temperature shift. A straight-line fit is shown as a guide to the eye. Figure adapted from [1].

6.7.4 Tolerance to Rate-of-Change of Temperature

The optical delay line was driven with triangular waveforms of different slopes to emulate different rates of temperature change from 0.01 to 0.45 °C/s. The impact of rate-of-change of temperature on clock phase caching for a series of different clock phase update rates is shown in Figure 6.16. Using the final rate-of-changes of temperature at which a CDR locking time under 625 ps is achieved in Figure 6.16, Figure 6.17 was generated. This figure shows that the critical rate of temperature change at which CDR locking time begins to increase to greater than under 16 symbols (under 625 ps) is proportional to the rate of clock phase updates. The proportionality constant, k, gives the clock phase update rate required to cope with the worst-case rate-of-change of temperature within the data-centre and is measured to be 27.0 $^{\circ}\mathrm{C}^{-1}$ in this experimental result.

This experimentally measured value of k is within 20% of the analytically modelled value of 33.6 $^{\circ}\mathrm{C}^{-1}$ calculated earlier in this chapter. Deviation between the experimentally measured and the analytically modelled values are likely to have arisen from the assumptions that were made in the analytical modelling to simplify its complexity, such as the pulse shape, which was assumed to be symmetric in the analytical modelling, but in practise was asymmetric in the experiment (as can be seen from the difference in the shape of the BER and CDR locking time profiles on the left and right side of Figure 6.13, which show the impact of positive and negative deviations of the clock phase sampling position from the optimal sampling point).

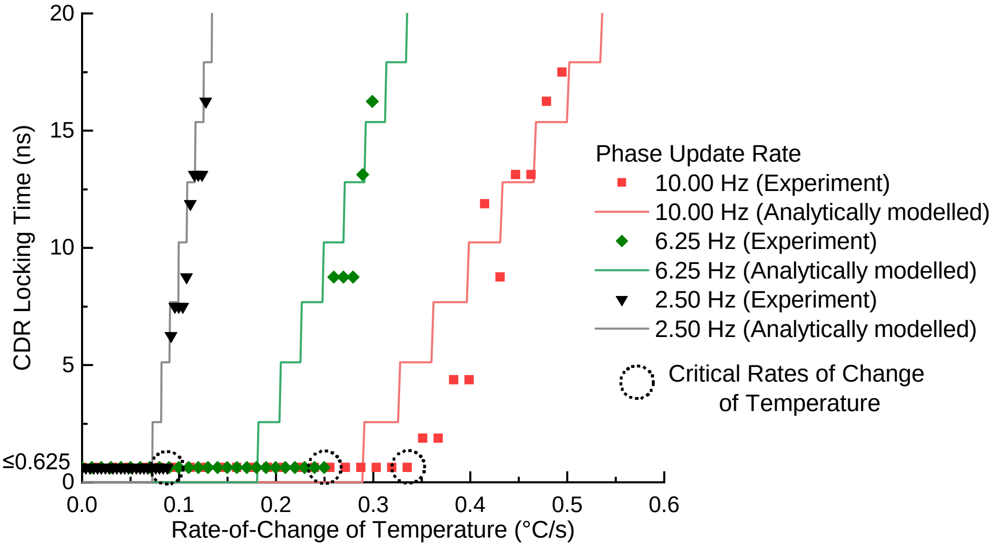

Fig. 6.16: Impact of rate-of-change of temperature on clock phase caching. The impact on CDR locking time of different rates of temperature and clock phase cache update rates across 2 km of SMF-28 clock fibre and 2 km of SMF-28 data fibre are shown (red, green and black points in the figure). The solid lines show the result predicted by analytical modelling in Figure 6.6 using the same parameters as those used experimentally. Figure adapted from [1].

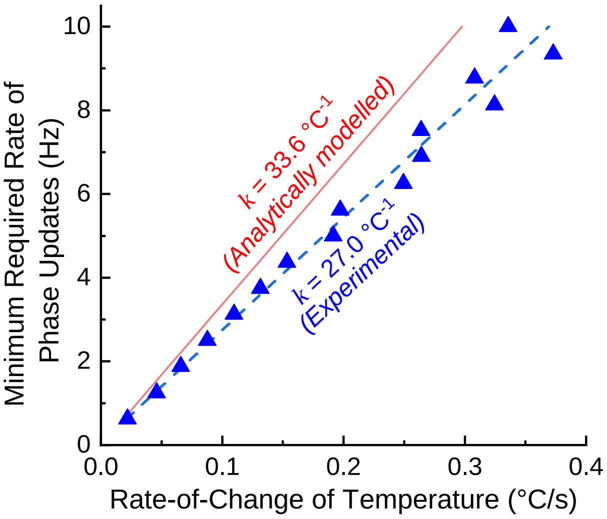

Fig. 6.17: Minimum required rate of clock phase updates to achieve under 625 ps (under 16 symbols) CDR locking time for different rates of temperature change across 2 km of SMF-28 clock fibre and 2 km of SMF-28 data fibre. Blue points: experimentally measured minimum required rates of clock phase update. Clock phase updates are required at a rate of 27.0 Hz for every 1 °C/s of rate-of-change of temperature (note that the worst-case rate of change of temperature was observed in a production cloud data centre was 0.03 °C/s). A linear regression fit (dashed blue line) was used to calculate this proportionality constant, which is within 20% of the analytically modelled value of 33.6 $°\mathrm{C}^{-1}$ (red solid line). Figure adapted from [1].

6.7.5 Tolerance to Source Clock Jitter

To investigate the jitter tolerance of clock phase caching, different amplitudes of 1 MHz sinusoidal jitter and white noise jitter were applied to the 800 MHz reference clock source. Jitter was applied to the electronic clock source shown in Figure 6.8 by frequency modulating the 800 MHz clock source with a 1 MHz sinusoidal and white noise voltage waveforms of different amplitudes. The peak-to-peak sinusoidal jitter applied to the clock was calculated after the clock output and before the MZM using an radio frequency spectrum analyser by measuring the ratio of the power of the 800 MHz clock tone to the power of its ±1 MHz side-tones. The root mean square (RMS) jitter amplitude of the clock after frequency modulation with white Gaussian noise jitter was measured using a digital communications analyser. 10^{11} bits per 16 bit bin were collected for each applied jitter amplitude. To prevent the receiver CDR remembering the clock phase of incoming data between successive packets, the receiver CDR phase was reset within the 5 ns inter-packet gap.

For both types of jitter, increasing the clock jitter amplitude eventually degrades the CDR locking time. The 2×2 km fibres were kept in an insulated chamber to minimise the impact of ambient temperature variation. The CDR locking time was measured against RMS jitter amplitude for 1 MHz sinusoidal jitter as shown in Figure 6.18 and white noise jitter as shown in Figure 6.19. The CDR locking time was under 625 ps (under 16 symbols) when the Gaussian jitter was equal to or less than 5.6 ps. In contrast, standard reference oscillators have jitter of a few hundred femtoseconds, which would ensure sufficient performance for clock phase caching for intra-data-centre interconnection.

Figure 6.7, shown in the analytical modelling section of this chapter, shows an analytical modelling analogue of the experimentally evaluated Figure 6.19. There is a significant difference in the point at which the CDR locking time begins to degrade from 0 or 0.625 ns to greater timescales: 0.145 symbols in Figure 6.19 and 0.075 symbols in Figure 6.7. However, the two results investigate the impact of different random jitter at different points in the clock phase cached system. In the experimental result, the *reference clock jitter* was varied. In the analytically modelled result, the *data clock jitter* was instead varied. The analytically modelled result does not account for the effect of the transmitter-side and receiver-side FPGA integrated inductor-capacitor tank (LC-Tank) quad phase locked loops (QPLLs) (as shown in Appendix B Figure B.1) that are used to multiply the 800 MHz reference clock to the half-rate 12.8 GHz serial clock that is input to the transmitter and receiver within the FPGAs. This LC-Tank PLL is likely acting to attenuate jitter present in the reference clock.

Evidence for this may be found in the large difference in the value of the total jitter measured from the data without injection of additional jitter into the reference clock (0.0604 symbols) versus the total jitter measured directly from the reference clock without additional injected jitter into the reference clock (0.0963 symbols). As a consequence of this, the experimentally evaluated Figure 6.19 and the analytically modelled Figure 6.7 are not comparable without adjusting the analytical model to model reference clock jitter rather than data jitter, which would require analytical modelling of the impact on clock jitter of the built-in LC-Tank PLL. This, in-turn, would require experimental characterisation of the LC-Tank PLL, as Xilinx has not made the jitter transfer characteristics of their LC-Tank PLL publicly available.

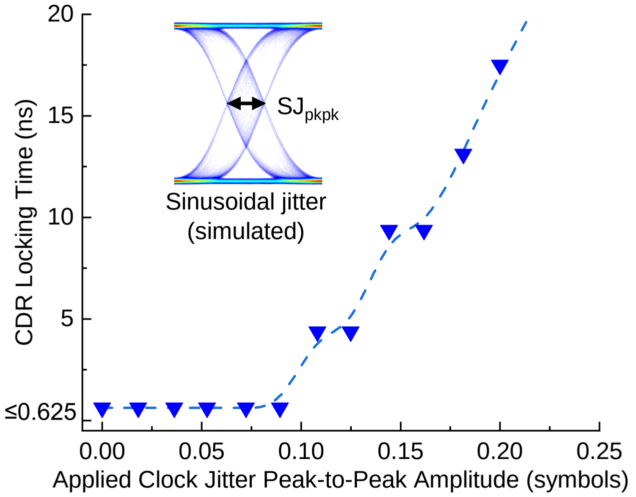

Fig. 6.18: Impact of 1 MHz sinusoidal jitter on CDR locking time. The tolerance to applied sinusoidal jitter is 0.090 symbols, corresponding to approximately 3.5 ps at 25.6 Gb/s. A B-spline fit is shown as a guide to the eye. Figure adapted from [1].

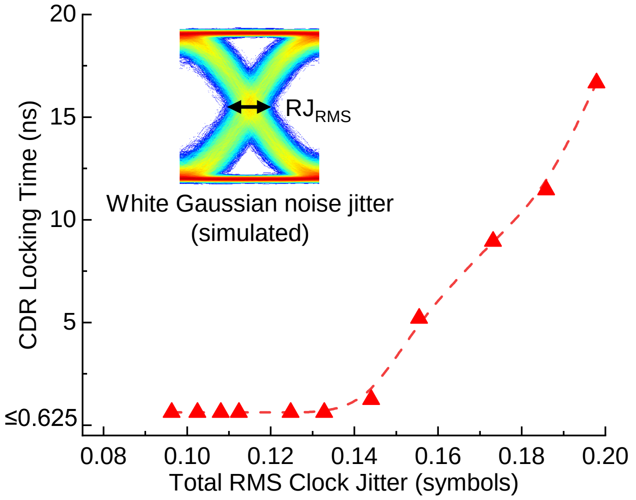

Fig. 6.19: Impact of white Gaussian noise jitter on CDR locking time. The tolerance to Gaussian jitter is 0.135 symbols, corresponding to approximately 5.2 ps RMS jitter at 25.6 Gb/s. A B-spline fit is shown as a guide to the eye. Figure adapted from [1].

6.7.6 Measurement of Clock Phased Cached Receiver Jitter

The tolerance of clock phase caching to additional clock jitter within the source clock, and to rate of change of temperature, is dependent on the relative clock jitter between the incoming data signal and the receiver sampling clock. Additionally, the clock phase caching algorithm must sample a sufficient number of incoming clock phase values to accurately measure the clock phase offset. To investigate this, two measurements of clock jitter were made. For both measurements, the experimental setup shown in Figure 6.8 was used, with the MZM 2×2 optical switch biased to transmit only packets from Node 0 to Node 2, with the 2 km of clock and 2 km of data fibre insulated from environmental temperature changes within the thermally controlled chamber.

In the first measurement, the clock phase offset values output from the Node 0 CDR were recorded over a 10 second period, with 2^{32} total clock phase values captured, with a clock phase measurement resolution of $\frac{1}{64}$ symbols. Clock phase caching was running with a clock phase update rate of 10 Hz The distribution of the recorded clock phase offset values is shown in Figure 6.20. The distribution is approximately Gaussian, with a single-term Gaussian fit resulting in a RMS standard deviation of 0.0210 symbols. This clock jitter arises from low-frequency jitter that is initially untracked by the CDR when packets first arrive, but is then tracked by the receiver CDR as it locks to the optimal sampling phase of incoming packets. This measured value for the low-frequency jitter standard deviation, $k_{\text{jit}-1}$, is used for the analytical modelling earlier in this chapter and in Chapter 5.

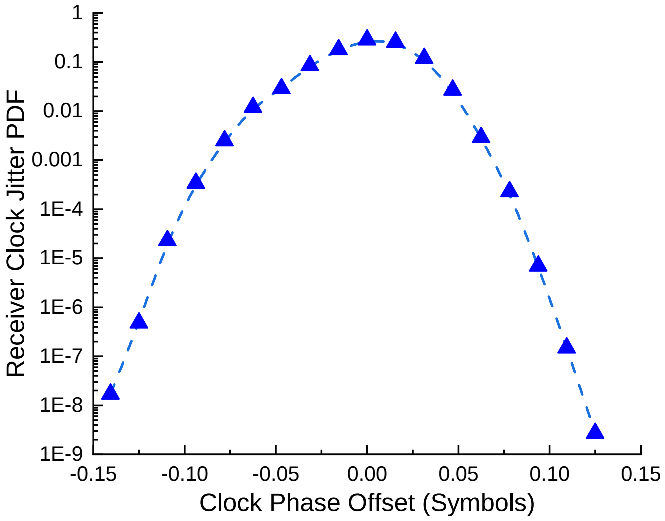

Fig. 6.20: Receiver CDR recorded clock phase offset value distribution. The distribution of these clock phase offsets is determined by the magnitude of the low-frequency jitter that is tracked by the CDR. A single-term Gaussian fit of the distribution results in a low-frequency random jitter standard deviation, $k_{\text{jit}-1}$, of 0.0210 symbols.

In the second measurement, the output of the Node 2 data PIN photodiode was brought into a digital communications analyser, which was triggered from the clock output of the Node 2 clock PIN photodiode. The total RMS jitter between the received data and the triggering clock was then measured using the digital communications analyser over a 1 minute period to be 2.36 ps (0.0604 symbols for 25.6 GBaud NRZ-OOK). This measured value was used for the total jitter standard deviation, k_{jit}, in the analytical modelling earlier in this chapter and in Chapters 4 and 5.

6.8 Estimated Scalability of Clock Phase Caching

Clock phase caching must be scalable to support large-port count data-centre optical switches. The worst-case throughput overhead resulting from clock phase caching can be estimated at different data-centre distance scales. An estimate of the worst-case optical switch throughput resulting overhead from clock phase caching at different distance scales, L, and number of servers and/or switches connected to the optical switch, N, can be found using Equation 6.15.

To obtain an estimate of throughput overhead from clock phase caching, the threshold maximum initial clock phase offset beyond which error degradation to worse than 10^{-10}, $\phi_{\max\,(\mathrm{thresh})}$, must first be calculated using Equation 6.11. Using the measured proportionality constant, k, of 27.0 °C, the experimental clock and data fibre length, L, of 2 km, an SMF-28 fibre thermal coefficient of delay, τ, of 42.6 ps/(km·°C) and a symbol rate, B, of 25.6 GBaud, $\phi_{\max\,(\mathrm{thresh})}$, gives an experimentally derived value for $\phi_{\max\,(\mathrm{thresh})}$ of 0.162 symbols.

An estimate of the worst-case optical switch throughput overhead from clock phase caching at different data centre scales can then be calculated using Equation 6.15 and the calculated value of $\phi_{\max\,(\text{thresh})}$. The following quantities will also be used: the experimentally measured value for k of 27.0 °C^{-1}; an SMF-28 thermal coefficient of delay, τ, of 42.6 ps/(km·°C); a symbol rate, B, of 25.6 GBaud; the measurement of worst-case rate-of-change of data-centre temperature, $\frac{dT}{dt}$, of 0.03 °C/s given earlier in this chapter; a time taken for each clock phase update per transmitter to receiver pair, t_{meas}, of 2.08 μs and a time taken to transmit each update packet, t_{update}, of 0.065 μs.

The resulting estimate of the worst-case optical switch throughput overhead from clock phase caching at different data centre scales is then shown in Figure 6.21. The estimated worst-case throughput overhead resulting from clock phase caching is only 1.7% for 10,000 data-centre end-points (servers or switches), all separated by 2 km. If 10,000 top-of-rack switches, each networking 64 servers (640,000 servers in total), were connected by an optically-switched fabric, clock phase caching would allow all possible transmitter to receiver pairs to communicate with each other with under 625 ps (under 16 symbols) CDR locking time with only a 1.7% optical switch worst-case throughput overhead. For 10,000 end-points, a traditional asynchronous CDR would need a clock recovery time of 25 symbols to have an equivalent throughput to clock phase caching.

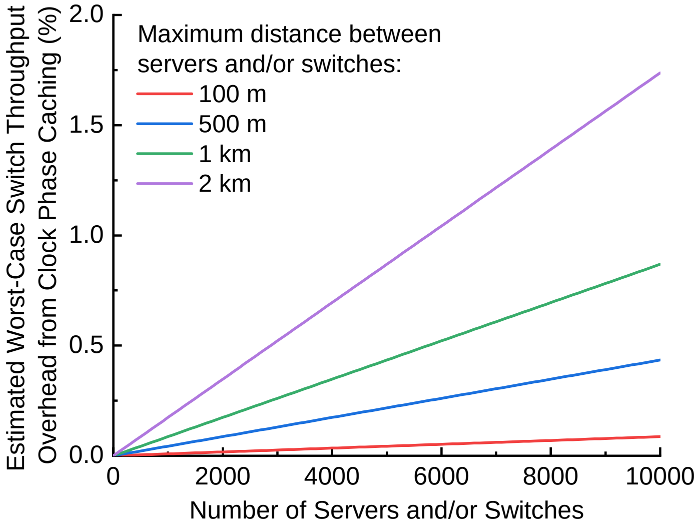

Fig. 6.21: Estimated worst-case optical switch overhead from clock phase caching. The overhead resulting from clock phase caching is directly proportional to the number of clock phase cached end-points connected to an optical switch. Clock phase caching supports all intra-data-centre distance scales with small throughput overhead from performing clock phase cache updates. At the largest intra-data-centre distance scale, with 10,000 servers or switches interconnected by an optical switch with a maximum of 2 km optical fibre between end-points, the estimated overhead from clock phase caching is only 1.7%. Figure adapted from [1].

Further improvements in scalability and/or reduction in overhead could be gained by reducing the rate of clock phase change for a given rate of temperature change across the same length of optical fibre. This could be achieved by using fibre with a smaller thermal coefficient of delay, such as HCF [21], which will be explored as part of Chapter 7. As an alternative approach, transmission between data-centre end-points through an optical switch can also be coordinated by a fixed cyclic schedule [8, 64], resulting in optical switch performance that matches that of an equivalent non-blocking electronically switched network [8]. In this case, clock phase caching would introduce no additional overhead as transmissions between Tx-Rx pairs are already forced by the scheduling algorithm.

6.9 Contribution Statement

The analytical modelling presented in this chapter was conceived, constructed and numerically evaluated by K.A.C., supervised by Z.L.. The experimental work presented in this chapter was originally published as a post-deadline conference paper at ECOC 2018 [19] and as a journal paper in Nature Electronics [1]. In these papers, the following contributions were made: K.A.C., P.C., H.B., I.H., K.J., B.T., H.W., P.W. and T.G. conceived the concept of clock phase caching, which was later refined with help from K.S., D.C. and Z.L.. K.A.C. and Z.L. conceived and constructed the experimental setup. K.A.C. implemented clock phase caching in the experiment. K.A.C. designed, implemented and tested all FPGA hardware code, with advice given by G.Z.. K.A.C. led the experiment and collected all experimental results, supervised by Z.L., with support provided by P.C., H.B., B.T., P.B. and G.Z.. P.C. and H.B. collected data-centre traffic and performed the optical switch network utilisation analysis (which is presented in this thesis in Chapter 1). K.A.C., Z.L., P.C. and H.B. wrote and revised the paper manuscript. All authors discussed the results and commented on the manuscript. The paper manuscript has been heavily altered for presentation in this thesis.

Chapter 7

Hollow Core Fibre Synchronisation for Sub-Nanosecond CDR Locking Time

7.1 Introduction

In Chapters 4 and 5, an approach to CSA-CDR was explored where the clock phase offset of all paths through an optical switch are calibrated on startup, followed by no subsequent phase updates. Chapter 4 explored the case where the receiver CDR circuits are not used to track the clock phase of incoming data packets, and Chapter 5 explored the case where the receiver CDR circuits do track the clock phase of incoming data packets. Irrespective of whether the CDR circuits are used, if there is only one calibration of the clock phase offset values, the large thermal coefficient of delay of SMF-28 would result in a very small temperature window for all but intra-rack distance scales, which would not be practical due to temperature variation in the data centre environment, which can be over 40 °C in the worst case.

A potential approach to rendering single calibration synchronisation viable is to use alternative optical fibres with a smaller thermal coefficient of delay to transmit the clock and data signals. This approach is explored in this chapter, with a focus on using HCF. This approach is first explored using the analytical model established in Chapters 4 and 5. The approach is then be experimentally confirmed in a real-time optically switched prototype. The temperature tolerance of a single calibration CSA-CDR without packet clock phase tracking system will first be studied in a point-to-point transmission experiment without any active clock phase updates following initial calibration. Then, the clock phase variation of HCF is measured and compared with SMF-28 in a 2×1 optically-switched system, in a demonstration of single calibration clock synchronisation CDR with packet clock phase tracking. The analytical modelling presented in this chapter is presented for the first time. The experimental work presented in this chapter was first published as a top-scored conference paper at ECOC 2019 [109] and was then published as an invited journal paper in the Journal of Lightwave Technology [2].

7.2 Using Low-TDC Fibre with Single Calibration CSA-CDR

CDR requires recovery of both the frequency and phase of the clock embedded in received data signals. The synchronisation of transceiver clock frequencies can be achieved by distributing an optical reference clock to all network nodes [83]. Nevertheless, as explored in Chapters 2, 5 and 6, clock phase recovery can still take up to 100s of nanoseconds because the CDR module must search through a range of clock phases for each new incoming data signal. This search for the correct sampling clock phase is necessary because of the change of the signal propagation delay due to environmental temperature variation. In a practical data centre environment, the recommended operating temperature range is from 18 to 27 °C [22] and the rate of temperature change in a typical production cloud data centre can be up to 0.03 °C/s [19]. For SMF-28, the typical rate of temperature-induced propagation delay change (given by the TCD) is about 40 ps/(km·°C) [21]. Assuming a typical 1 km short-reach data centre link, a 0.5 °C temperature change would lead to a data delay change of half a symbol period for 25 GBd signals, causing bit errors if a single calibration synchronisation assisted approach to CDR is used, as explored in Chapters 4 and 5. Chapter 6 established clock phase caching, which allows this clock phase change due to temperature shift to be compensated for by performing regular clock phase updates, providing a solution to this problem.

A possible alternative / complementary solution, first explored using analytical modelling in Chapters 4 and 5, could be to calibrate the clock phase values only once. In such a system, an optically-switched network could be constructed that uses fixed clock phase offsets, calibrated once on network startup, which thereafter uses static clock phase values without any need for clock phase updates as the data centre temperature changes. This single calibration approach to CSA-CDR could eliminate the need for active clock phase updates used in clock phase caching, avoiding the network overhead associated with this, which, though it is small, 1.7% for 10,000 nodes with a maximum of 2 km of fibre between end-points, it nonetheless has an impact. As explored in Chapters 4 and 5, for SMF-28 fibre, the temperature range over which this is possible, less than 1°C for a cluster-scale optically-switched network, is too small to be practical.

Low-thermal sensitivity optical fibres could potentially greatly widen the allowable temperature range of an optically-switched network that uses a static calibration approach to CSA-CDR. Additionally, low-thermal sensitivity optical fibre could be used to reduce the required rate of clock phase updates in a clock phase cached system, thereby reducing the network overhead. Although low thermal sensitivity liquid crystal polymer coated SMF-28 is available with a reduced temperature delay coefficient (TDC) of under 5 ps/(km·°C) [94], the largest reductions in TDC are offered by HCF, which have been demonstrated to have a TCD of 2 ps/(km·°C), which is about 20 times lower than the TCD of SMF-28 [21]. A specific design of HCF has also been demonstrated to have a TCD equal to 0 ps/(km·°C) [110]. Although this was achieved only at a specific wavelength, it potentially opens new opportunities for ultra-fast CDR that is not affected by the impact of environmental temperature variation.

An important characteristic warranting consideration for HCF is optical attenuation, which was previously much greater than in SMF-28. The latest generation of HCF has already shown a loss below 1 dB/km (recently-published record is 0.65 dB/km at 1550 nm [111]) with further loss reduction anticipated. They are also now manufactured in kilometre lengths [112]. Furthermore, they can have low chromatic dispersion (e.g. $>$3 ps (nm·km)) and very low non-linearity ($>$2 orders of magnitude lower than SMF-28). They also offer shorter propagation delay (33% lower latency), as signals propagate about 1.46 times faster in HCFs as compared to SMF-28 [113]. All these features and the latest fabrication developments makes them particularly interesting for data centre applications.

7.3 Thermal Properties of Hollow Core Fibre

As first explored in Chapter 4, signal propagation time through an optical fibre depends on temperature. This dependence arises from two additive effects: the change of the refractive index of silica glass with temperature, which causes the light travels more slowly through the fibre as the temperature is increased, and fibre elongation due to the thermal expansion of the silica glass with temperature, which causes the light to travel a longer physical distance as the temperature increases. The change in fibre time-of-flight as temperature increases, $\frac{dt}{dT}$, as first given in Chapter 4, is given by [21]:

$$\frac{dt}{dT} = \frac{1}{c}\left(n_g\frac{dL}{dT} + L\frac{dn_g}{dT}\right) \tag{7.1}$$

where c is the speed of light in vacuum, n_g is the group refractive index of the fibre and L is the fibre length.

The first term in Equation 7.1 gives the effect of fibre length elongation. The second term in Equation 7.1 gives the effect of variation in group refractive index of the optical confinement medium of the optical fibre. The optical confinement method and medium in HCF are different to that used in standard SMF-28. In SMF-28, the transmitted optical signal is confined to a central doped glass core by total internal reflection. In HCF, the transmitted optical signal is confined to a central hollow air or vacuum filled core by forbidding propagation modes outside the core by using a repeating lattice structure to cause destructive interference of those modes [114].

For SMF-28, the dominant effect is the change of the refractive index of the silica glass core with temperature, contributes 37 ps/(km·°C) to the total thermal sensitivity. The other effect, fibre elongation due to the thermal expansion of the silica glass core with temperature, contributes 2 ps/(km·°C) to the total thermal sensitivity. This leads to an overall theoretical thermal sensitivity of approximately 39 ps/(km·°C), which closely matches experimental measurements of thermal sensitivity [94]. For HCF, the refractive index change of the air or vacuum within the fibre core is much smaller than that of silica, providing a much smaller contribution to the thermal sensitivity. For hollow-core photonic bandgap fibre (HC-PBGF) (as used in this chapter), this effect contributes approximately 0.13 ps/(km·°C) to the fibre thermal sensitivity [21]. The contribution from fibre elongation is also lower in HCF than in SMF-28. For the HC-PBGF, this effect contributes approximately 1.42 ps/(km·°C) to the fibre thermal sensitivity [21]. This gives an overall theoretical thermal sensitivity for the HC-PBGF of 1.6 ps/(km·°C), approximately 20× smaller than that of SMF-28.

Figure 7.1 shows a measurement of the delay change versus temperature for a spool of 1 km of SMF-28 and likewise for a 1 km length of HC-PBGF. The propagation delay for the measured SMF-28 spool changed at a rate of 46 ps/°C, corresponding to a TCD of 46 ps/(km·°C). The much smaller thermal sensitivity of HCF was confirmed by the HC-PBGF propagation delay with temperature measurements in Figure 7.1, which shows a drift of 1.8 ps/°C, corresponding to a TCD of 1.8 ps/(km·°C).

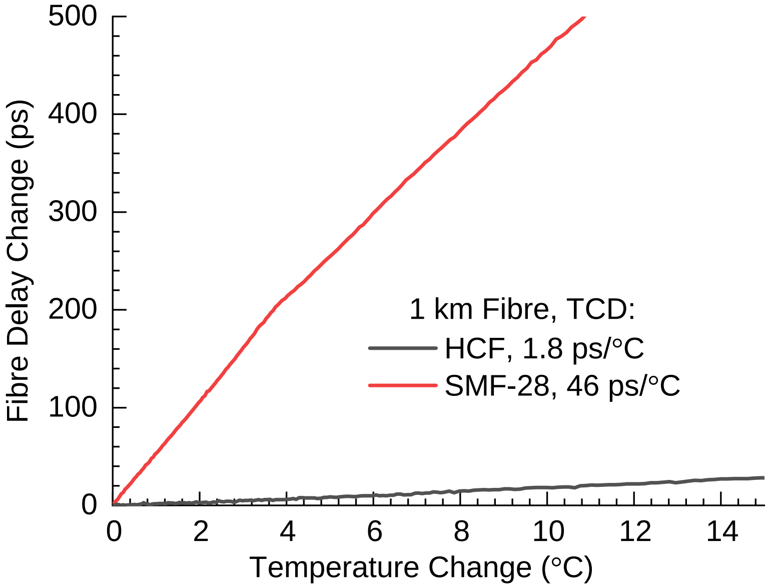

Fig. 7.1: Measured change of delay versus temperature change for a 1550 nm signal propagating through 1 km lengths of SMF-28 (red line) and HCF (black line). The thermal coefficient of delay of HCF is approximately 20 times smaller than that of SMF-28. Figure adapted from [2].

7.4 Analytical Modelling of Single Calibration CSA-CDR with Hollow Core Fibre Transmission

This section will use the analytical modelling of single calibration CSA-CDR without packet clock phase tracking introduced in Chapter 4 to assess power penalty of that approach with HCF transmission, and will use the analytical modelling of single calibration CSA-CDR with packet clock phase tracking introduced in Chapter 5 to assess CDR locking time. The same assumptions and parameters used in Chapters 4 and 5 will be used to model these performance measures, with the exception of the fibre thermal sensitivity of delay, τ, for which a value of 1.8 ps/(km·°C) will be used, matching the experimentally measured value in the previous section.

The proportionally smaller value of the fibre TCD of HCF versus SMF-28 reduces the sampling clock phase change experienced by the clock and data transmission fibres for the same change in temperature. The relationship between fibre TCD, τ, the distance between nodes, L, the data symbol rate, B, and the change in temperature, ΔT, accounting for both clock and data fibre contributing to the clock phase shift, is:

$$\Delta\phi = 2\tau LB\Delta T \tag{7.2}$$

7.4.1 Optical Power Penalty from Single Calibration CSA-CDR Without Packet Clock Phase Tracking

Figure 7.2 shows the power penalty from single calibration CSA-CDR without packet clock phase tracking, with HCF clock and data transmission, as a function of temperature change, at each of the four typical data centre length scales (Figure 7.2a, intra-rack ($\leq$7 m); Figure 7.2b, intra-cluster ($\leq$100 m); Figure 7.2c, core ($\leq$2 km) and Figure 7.2d inter-building ($\leq$10 km)), for a selection of different jitter standard deviations.

Assuming a total random jitter of a Xilinx UltraScale GTY 25 Gb/s transceiver of 2.08 ps (as measured from a Xilinx UltraScale GTY transceiver in Chapter 6), a jitter standard deviation of 2 ps could therefore be considered reasonable for a clock synchronised 25 GBaud NRZ FPGA transmitter. If a 1 dB power penalty resulting from clock phase offset from operating without the CDR is also considered reasonable, then the predicted range of acceptable temperatures for each length scale, with HCF transmission, are 224 °C for intra-rack ($\leq$7 m), 15.7 °C for intra-cluster ($\leq$100 m), 0.783 °C for core ($\leq$2 km) and 0.157 °C for inter-building ($\leq$10 km).

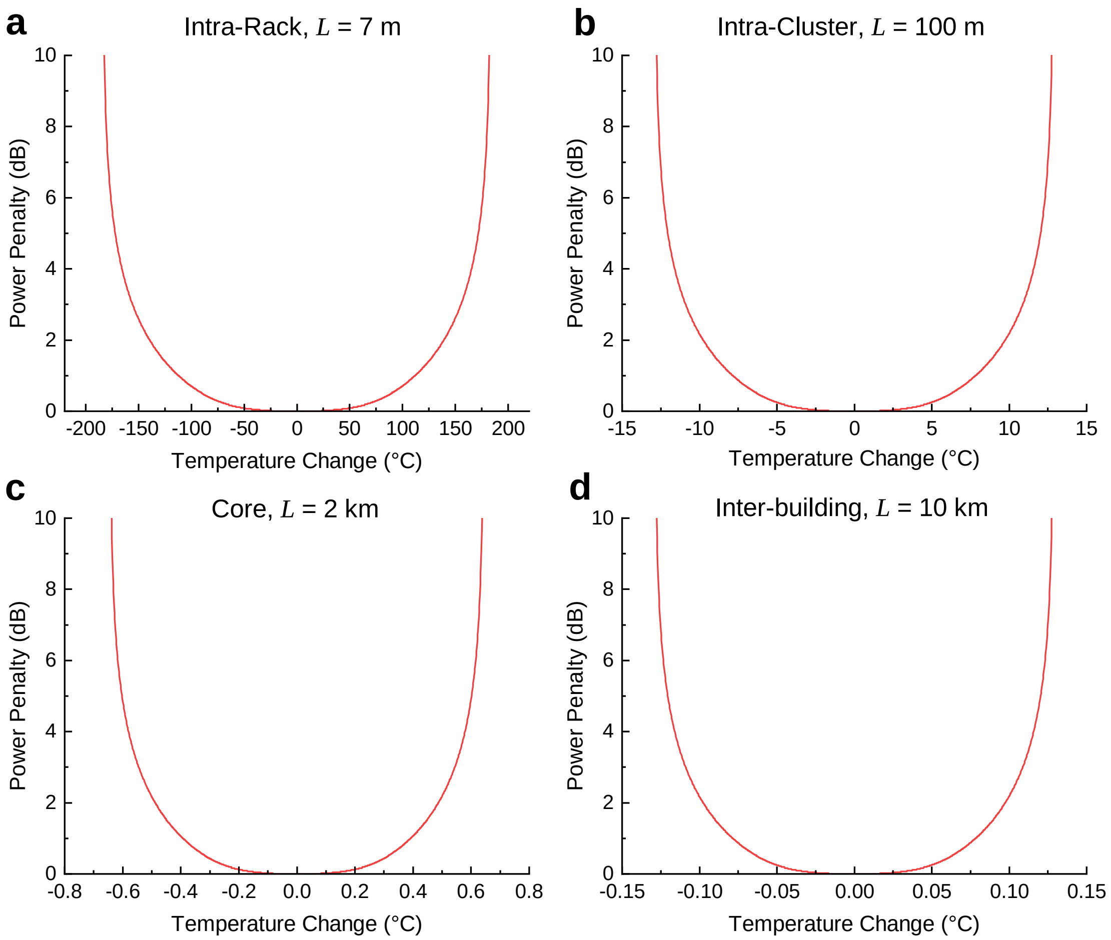

Fig. 7.2: Analytically modelled power penalty to maintain a bit error probability of 10^{-10} at a receiver, resulting from temperature change in an optically-switched network, operating with single calibration CSA-CDR without packet clock phase tracking, with HCF clock and data transmission. The power penalty is modelled for four different jitter magnitudes, at four different data centre length scales. NRZ-OOK signal reception is modelled with a symbol rate of 25 GBaud using typical data centre TIA and PIN photodiode characteristics. **a**, intra-rack ($\leq$7 m); **b**, intra-cluster ($\leq$100 m); **c**, core ($\leq$2 km); **d**, inter-building ($\leq$10 km).

7.4.2 CDR Locking Time using Single Calibration CSA-CDR With Packet Clock Phase Tracking

Figure 7.3 then shows the CDR locking time as a function of temperature change for a Tx-Rx link operating with single calibration CSA-CDR without packet clock phase tracking, with HCF used for clock and data transmission, at each of the four typical data centre length scales (Figure 7.3a, intra-rack ($\leq$7 m); Figure 7.3b, intra-cluster ($\leq$100 m); Figure 7.3c, core ($\leq$2 km) and Figure 7.3d, inter-building ($\leq$10 km)), including the impact of high and low-frequency jitter for a typical synchronised data centre FPGA transceiver. The predicted range of temperatures for which the CDR locking time is instantaneous are 324 °C for intra-rack ($\leq$7 m), 22.7 °C for intra-cluster ($\leq$100 m), 1.13 °C for core ($\leq$2 km) and 0.227 °C for inter-building ($\leq$10 km).

7.4.3 Theoretical Viability of Single Calibration CSA-CDR Approaches using Hollow Core Fibre Transmission

For single calibration CSA-CDR without packet clock phase tracking, the use of HCF for both clock and data transmission increases the temperature range over which under a 1 dB power penalty is achievable by approximately 20×. For single calibration CSA-CDR with packet clock phase tracking, the use of HCF for both clock and data transmission increases the temperature range over which instantaneous CDR lock occurs by approximately 20×. These temperature range increases arise from the approximately 20× smaller TCD of HCF versus SMF-28.

As first discussed in Chapter 2, the air temperature in a data centre designed according to standard recommendations may vary by up to 40 °C [22]. The A1 class of allowable temperature further restrict this to a 15 °C range respectively. The maximum temperature range for single calibration CSA-CDR without packet clock phase tracking with a 1 dB power penalty is greater than this, but the maximum allowable temperature range of single calibration CSA-CDR is under this temperature range for the intra-rack and intra-cluster / intra-cluster distance range. Both approaches are therefore likely to be viable for intra-rack and intra-cluster transmission. In contrast, if SMF-28 transmission is used, both approaches to single calibration CSA-CDR were predicted to only be viable for intra-rack communication, with clock phase caching is required at all other distance scales.

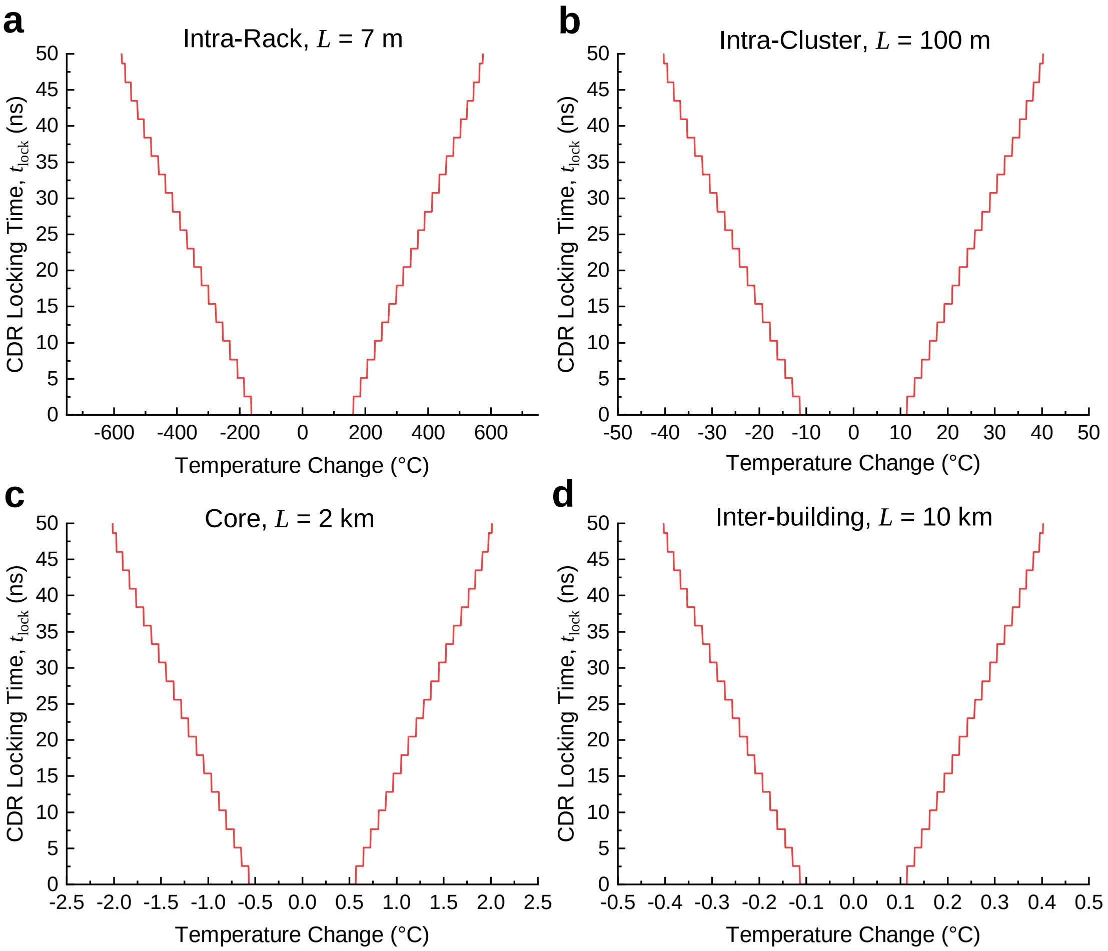

Fig. 7.3: CDR locking time against temperature change since clock phase calibration, with HCF clock and data transmission. These plots were modelled including the impact of high and low frequency jitter. A value of $k_{\text{jit}-\text{h}}$ of 0.566 symbols and a value of $k_{\text{jit}-\text{l}}$ of 0.0210 symbols were used to generate the CDR locking time values. A CDR error probability threshold of 10^{-10} was used to determine the CDR locking time. **a**, intra-rack ($\leq$7 m); **b**, intra-cluster ($\leq$100 m); **c**, core ($\leq$2 km); **d**, inter-building ($\leq$10 km).

7.5 Experimental Investigation of Single Calibration CSA-CDR Techniques with Hollow Core Fibre

An experimental investigation of these two single calibration CSA-CDR approaches were then performed to assess their performance and practical viability under temperature variation.

7.5.1 Experimental Setup

The effect of the low thermal sensitivity of HCF on single calibration CSA-CDR without packet clock phase tracking in a point-to-point transmission experiment is shown in Figure 7.4. An optical clock signal was generated by modulating an 800 MHz reference clock (Si5340, 320 fs rms jitter (50 kHz-80 MHz) [115]) onto a 1550 nm optical carrier, via a $LiNbO_3$ MZM. The optical clock signal was split and sent directly to the receiver (Rx FPGA, Xilinx VCU108) and to the transmitter (Tx FPGA, Xilinx VCU108) via a fully connectorised (FC/APC SMF-28 pigtails) 1 km HCF. The optical clock signals were detected and amplified for frequency synchronisation of the two FPGAs through digital PLL modules integrated within each FPGA. At the transmitter side, an EML consisting of a distributed feedback (DFB) laser outputting 13 dBm of CW power at 1555 nm followed by a 35-GHz EAM which was driven with 25.6 Gb/s NRZ PRBS data with a length of 2^{31}-1 bits [108]. The resulting 3-dBm output signal was subsequently launched into a second HCF of 1 km length. The end-to-end loss of the data and clock paths were 6.5 dB and 7.0 dB, respectively, consisting of about 3 dB connector loss and 4 dB fibre loss. At the receiver side, the signal was detected by a 20 GHz bandwidth photodiode followed by a TIA before the real-time FPGA receiver. The received optical data and optical clock powers were intentionally attenuated to -10.5 dBm to match minimum tolerable power for optical receivers in intra-data centre transmission standards [10]. Bit errors were counted in real-time and used to calculate BER. Note that the counter-propagation of the data and the clock signals in this experiment emulated the worst-case scenario in data centre networks, where the clock source is adjacent to the receiver and far from the transmitter, causing an additive change of the clock phase drift in the clock and data paths due to temperature change (as discussed in more detail in Chapter 4). Further details of the FPGA implementation of CSA-CDR are given in Appendix B. Two photographs of the experimental setup are shown in Appendix C.

To investigate the impact of clock phase offset due to temperature change on BER, both HCF spools were placed inside a thermally controlled chamber, in which the air temperature was monitored by six temperature sensors surrounding the fibre spools. After an initial period of stabilisation, which took about 5 minutes, the temperature fluctuated by less than 0.05 °C. The chamber was initially stabilised at 33.5 °C and

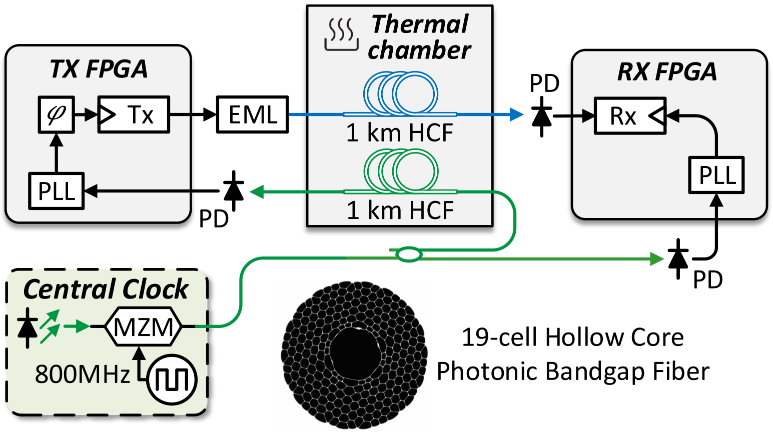

Fig. 7.4: Synchronous HC-PBGF point-to-point transmission experiment to investigate single calibration CSA-CDR without packet clock phase tracking. PLL, Phase locked loop; EML, externally modulated laser; MZM, Mach Zehnder Modulator; HCF, hollow core fibre; PD, photodiode. Figure reproduced from [2].

the receiver data sampling clock phase was set to be half a symbol period away from the optimum value. The optimum clock phase value was determined by measuring the Q factor of the received signal. The initial temperature of 33.5 °C was chosen to ensure a significant offset from ambient temperature, maximising the thermal chamber temperature stability at each point. The chamber temperature was then increased and the BER was recorded after the temperature stabilised (to within ±0.05 °C. Each BER point was measured from 7×10^{12} bits.

To evaluate the effect of the low thermal sensitivity of HCF on single calibration CSA-CDR with packet clock phase tracking, the effect of temperature change on the BER and CDR locking time was investigated in an optically-switched prototype system consisting of a 2×1 optical switch, two transmitters (Tx0 and Tx1) and a receiver, Rx, as shown in Figure 7.5. In this experiment, the optical clock was sent to Tx0 through 1 km of HCF, and directly to Tx1 and the Rx. Each packet contained a 128-byte payload followed by 64-byte control information. The payload was formed of two de Bruijn sequences that had a length of 2^9 bits (an identical sequence to the one used in Chapter 6 was used in this context), which was chosen to emulate the typical packet size of data centre traffic [10]. The optical packets were generated by driving EML0 and EML1 with the 25.6 Gb/s NRZ-OOK produced by Tx0 and Tx1. Their output optical signals were at 1555 nm and 1553 nm, respectively. The optical packets from both transmitters were of 60 ns duration (40 ns payload followed by 20 ns control information) and were interleaved by a 2×2 $LiNbO_3$ MZM with a 5 ns gap between the consecutive packets. The 2×2 MZM had a rise time of less than 1 ns and was used to emulate a fast optical switch. The interleaved packets leaving the switch were amplified by an EDFA to compensate for the loss of the 2×2 $LiNbO_3$ MZM (about 6 dB). The use of an EDFA would be unnecessary in a practical system due to recent improvements in HCF optical attenuation to below 1 dB/km [111]. Scalable and low loss fast optical switches can be used to provide sufficient power for optically-switched networks [13]. The interleaved packets then passed through 1 km of HCF before being received by the Rx node. In the same fashion as in the point-to-point experiment, the received optical power was attenuated to -10.5 dBm so as to be equal to the minimum required power in the Ethernet standard [10]. Both HCF fibre spools were contained within the thermally controlled chamber which was initially stabilised at 32.5 °C. The BER and the CDR locking time was measured at different temperature values. In this experiment, the receiver CDR module was kept running constantly. As a result, only the clock phase offset between the two transmitters contributed to the BER and CDR locking time, which is due to the temperature induced clock and data phase offset through the HCFs.

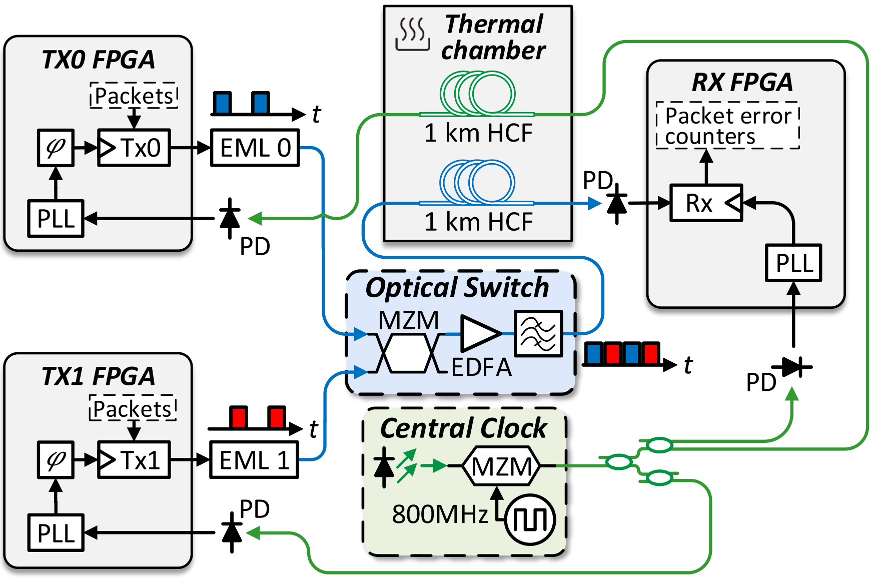

Fig. 7.5: Synchronous HC-PBGF 2×1 optical switching experimental setup to investigate single calibration CSA-CDR with packet clock phase tracking. EML, externally modulated laser; EDFA, erbium doped fibre amplifier. Figure reproduced from [2].

The overall BER and CDR locking time were calculated in the same fashion as in Chapter 6. To do this, the data payload was divided into 64 bins, each containing 16 bits (625 ps). The packets were detected in real-time by the FPGA receiver which recorded the number of bit errors falling in each bin until 10^{13} total bits (summed across all bins) had been received. The CDR locking time was calculated as the first packet bin with a BER of under 10^{-10}, if all following bins also had a BER of under 10^{-10}. The overall receiver BER was calculated by summing the errors falling in all 16-bit bins (a total of 64×10^{13} bits).

7.5.2 Experimental Results

Figure 7.6 shows the impact of temperature change on the BER of the point-to-point clock-synchronised transmission system used to investigate single calibration CSA-CDR without packet clock phase tracking. The blue markers show the measured BERs using the HCFs. The inset shows the eye diagram captured from the receiver FPGA. As the temperature increases, the counter propagated clock path and data path lead to a total receiver clock phase shift of about 4 ps/°C (equivalent HCF length of 2 km). Error free (BER under 10^{-12}) performance was observed over a temperature range of 2 °C (35 to 37 °C. To compare these HCF results with their SMF-28 equivalent, a thermal coefficient of 40 ps/(km·°C) was assumed, which was used with Equation 7.2 to estimate the tolerance to temperature variation based on the measured CDR locking time at different clock phase offset values shown in Chapter 6 Figure 6.5. The experimental setup used to acquire Chapter 6 Figure 6.5 is almost identical to the one used in the experiment in this chapter, with the only differences being the use of an optical frequency comb for the clock source in Chapter 6, and different optical fibre. Shown as cross markers in Figure 7.6, the error free temperature range using SMF-28 was only about 0.1 °C, which in practice would require clock phase updates. This was not directly measured experimentally using SMF-28 due to the precision of the fibre temperature required to record a reasonable measurement, and because the impact of temperature on time-of-flight change of SMF-28 is very well established in literature.

Figure 7.7a and 7.7b show the BER and CDR locking time, respectively, of the 2×1 optical switching system experiment used to investigate single calibration CSA-CDR with packet clock phase tracking. The CDR locking time was calculated from the BER values output from the FPGA receiver. In contrast to Chapter 6, in this experiment, the clock phase of the transmitters was not actively updated, i.e. no phase caching was employed to track the clock phase change with temperature. This allowed the performance using the two different fibre types to be directly compared. Using the HCF, a window of 4 °C temperature change was observed over which the received data had BER of under 10^{-12} and a CDR locking time of under 625 ps (under 16 symbols). The tolerance to temperature variation is consistent with the clock phase offset caused by the 1 km HCF length difference between the two clock-to-receiver-via-transmitter paths, as shown in the experimental setup in Figure 7.7. This tolerance to temperature variation of double that of Figure 7.6 is consistent with the clock phase offset between the transmitters only arising from the 1 km of clock fiber between the clock source and transmitter 0 (as shown in the experimental setup in Figure 7.5), which is half the length of fiber contributing to the clock phase change in Figure 7.6 of 2 km. This difference also causes the error-free temperature window to be centred on a higher temperature in Figure 7.7 than Figure 7.6. The equivalent temperature range for SMF-28 in Figure 7.7 was 0.2 °C, calculated in the same manner as Figure 7.6 but with 1 km fiber length.

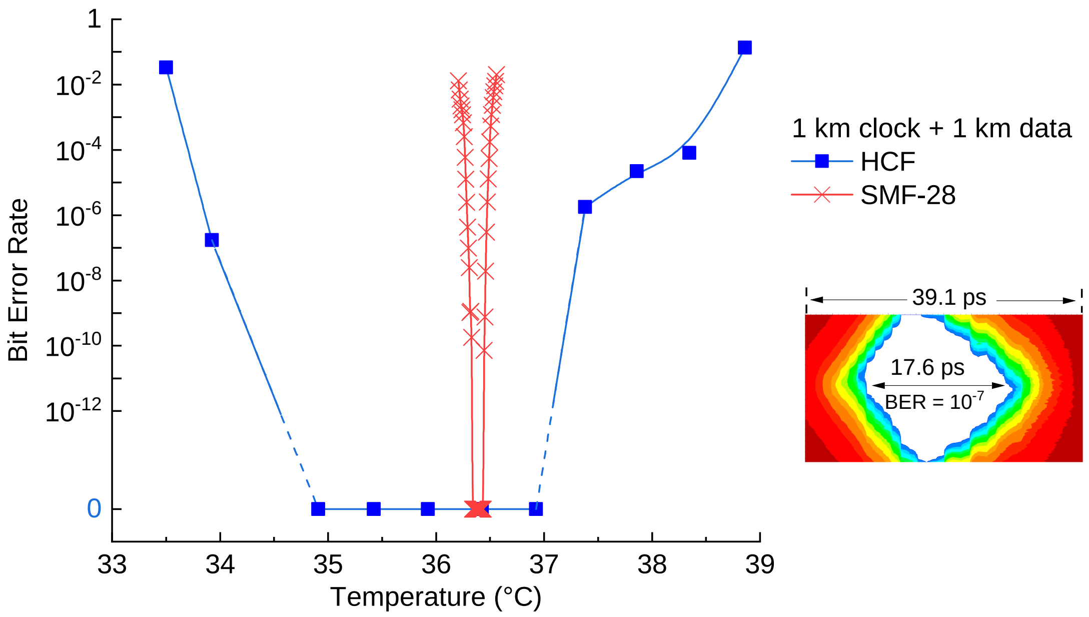

Fig. 7.6: Impact of temperature variation on the performance of a synchronous HC-PBGF point-to-point transmission system with single calibration CSA-CDR with packet clock phase tracking. Blue markers: measured BER using the HCF; Red cross markers: estimated from SMF-28. A B-spline fit is shown as a guide to the eye. Figure adapted from [2].

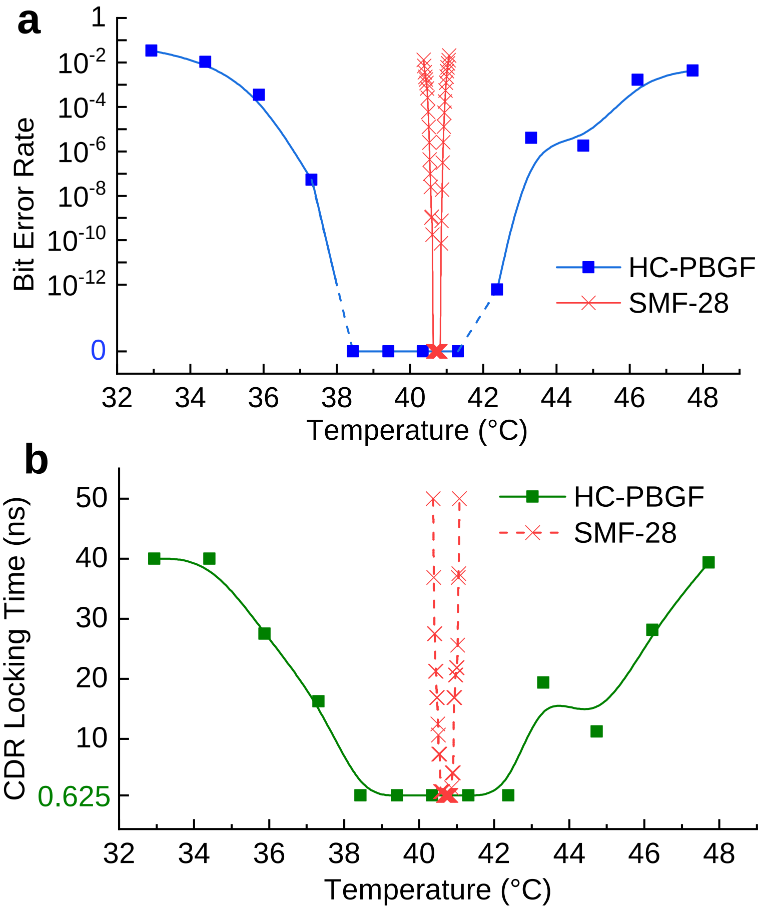

Fig. 7.7: Impact of temperature variation on the performance of a synchronous HC-PBGF optically-switched system with single calibration clock single calibration CSA-CDR with clock phase tracking. a, BER and **b**, CDR locking time. Blue markers: measured BER using HCF; Red cross markers: estimated for SMF-28. Green: CDR locking time. A B-spline fit is shown as a guide to the eye. Figure adapted from [2].

7.6 Discussion

These results confirm that the use of HCF should allow for short-reach data centre interconnection (e.g. less than 2 km fibre length) without the active phase updates that are required for SMF-28 based clock-synchronised networks. As discussed in Chapter 4, the change of time-of-flight through a fibre is linearly proportional to both the change in fibre temperature and the fibre length [21]. Therefore, based the experimental results from the previous section, i.e. under 625 ps (under 16 symbol) CDR locking time over a temperature range of 2 °C for 1 km HCF data fibre and 1 km HCF clock fibre, HCF based optically-switched systems would be able to support under 625 ps (under 16 symbol) CDR locking time over a 20 °C temperature tolerance for data fibre and clock fibre lengths of up to 100 m each.

Therefore, based on the experimental results, i.e. under 625 ps (under 16 symbol) CDR locking time over a temperature range of 2 °C for 2 km of HCF, HCF based optically-switched systems would be able to support a 20 °C temperature range for fibre lengths of up to 100 m. This confirms the analytical model prediction of single calibration synchronisation CDR being viable for this distance scale, which predicted an allowable temperature range of 22.7 °C. This would provide sufficient temperature tolerance [22] for connecting servers within intra-cluster or data centre cluster distance scales with only an initial calibration of the clock phase values, as long as the operational temperature is restricted to the ASHRAE A1 class (15 to 32 °C). Ideally, the allowable temperature windows of shorter lengths of HCF would have been measured directly, but only 1 km fibre lengths were available for this experiment.

Theoretically, larger temperature ranges at a given fibre length are possible. The upper bound on the under 625 ps (under 16 symbol) CDR locking time temperature window, $\Delta T_{\max}$, is given by:

$$\Delta T_{\max} = \frac{1}{\tau L B} \tag{7.3}$$

In the case of Figure 7.6 and Figure 7.7, with contributing fibre lengths of 2 km and 1 km respectively, and τ of 1.8 ps/(km·°C), $\Delta T_{\max}$ is therefore 10.9 °C and 21.7 °C respectively, versus the 2 °C and 4 °C ranges measured in this experiment. The significant difference between the measured values and these theoretical maximums arises from the influence of clock jitter and noise, which in practice reduces the acceptable range of clock phases offsets that can be tolerated.

The low thermal sensitivity of HC-PBGF can also be used to further improve the scalability and/or reduce the transmission overhead of a optically-switched system that uses clock phase caching[†]. The small thermal coefficient of delay of HC-PBGF reduces the rate of clock phase change that occurs for a given rate of temperature change, in turn reducing the rate at which clock phase updates are required to maintain clock phase synchronisation. The 20 times lower thermal coefficient of delay of HC-PBGF versus SMF-28 leads to a 20 times reduction in the rate of required clock phase updates for a given rate of temperature change. This, in-turn leads to an estimated overhead reduction by a factor of 20 (as shown in Figure 7.8a), scaling up of the number of supported nodes at the same overhead by a factor of 20 (as shown in Figure 7.8b), or a combination of these. Figure 7.8a and b were calculated using Chapter 6 Equation 6.15 and the parameters used in the calculation of Chapter 6 Figure 6.21, with the exception of the fibre thermal coefficient of delay, for which the 20 times lower thermal coefficient of delay of HC-PBGF of 1.8 ps/(km·°C) was used.

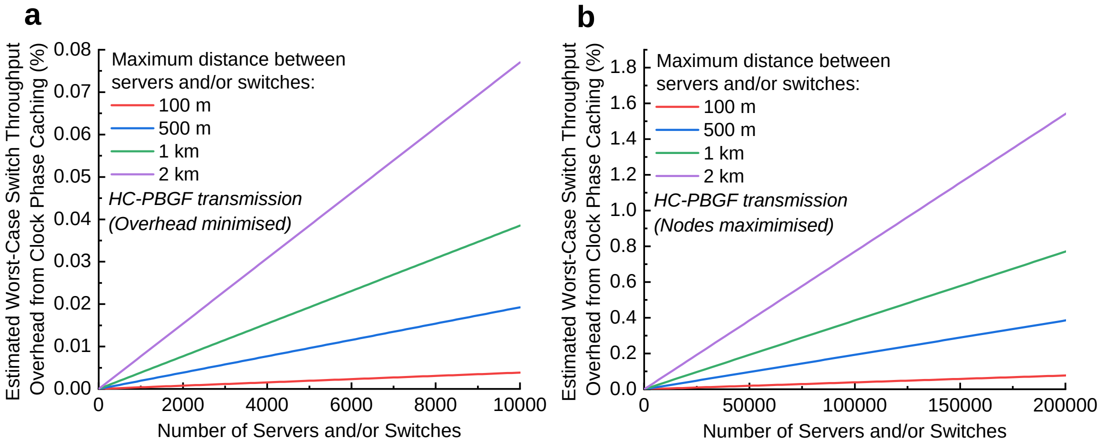

Fig. 7.8: Estimated worst-case optical switch overhead from clock phase caching, with low-thermal sensitivity HCF transmission. The overhead resulting from clock phase caching is directly proportional to the number of clock phase cached end-points connected to an optical switch. The low thermal sensitivity of HCF can be used to minimise the worst-case optical switch overhead from the clock phase caching technique to negligible levels (shown in **a**), maximise the number of nodes that can be interconnected by a single optical switch with low clock phase caching overhead (shown in **b**), or a combination of these. Figure adapted from [1].

[†] Note that even with SMF-28 transmission, a clock phase cached system is still already highly scalable (with 10,000 servers and/or switches leading to only 1.7% optical switch transmission overhead under worst-case data centre environmental conditions), but HC-PBGF can nonetheless result in a 20 times further improvement in scalability.

HC-PBGF does have multi-mode properties, but it can be effectively single modal with optimal launch conditions, as shown in [116]. Therefore, with optimised SMF-28 to HCF launch, the mode partition penalty is negligible as shown in [114, 117] for 250 m HC-PBGF and so is not a concern for clock synchronisation. The overall connectorisation loss of 2.5 dB and the attenuation of 4.5 dB/km at 1553 nm of the HC-PBGFs used in the experiment is greater than that of SMF-28 (typically 0.25 to 0.5 dB per connectorisation and 0.17 dB/km attenuation respectively). This 7 dB loss is still smaller than the power budget of intra-data center interconnection [10]. However, the 4.5 dB/km attenuation of HC-PBGF could be an concern for high-radix centralised clock distribution or high-radix optical switching over 1 km data center distance scales, where the loss would reduce the possible radix versus SMF-28 fibre based clock distribution. Low loss fiber is also preferred for low power interconnects and improved resilience.

Recent advances in HCF technology have achieved large reductions in loss. The 3 dB overall connecterisation loss in the experiment, resulting from the mode field diameter (MFD) mismatch between SMF-28 and HC-PBGF, can be reduced to 0.3 dB per connectorisation [118]. In addition, the latest generation of HCF, hollow-core nested anti-resonant nodeless fiber (HC-NANF), has already shown an attenuation below 1 dB/km. The most recently published record is 0.65 dB/km at 1550 nm [111], with further loss reduction anticipated. HC-NANFs are also now manufactured in kilometre lengths [119]. The structure and therefore guiding mechanism of these fibres differs from the HC-PBGF used in the experiment in this chapter, but in both types, 99% of the light is guided through the hollow air core as described earlier in this chapter, yielding a thermal coefficient of delay of approximately 2 ps/(km·°C), similar to that of HC-PBGF [120].

The results in this chapter show that clock-synchronised data center systems can have 20 times greater tolerance to temperature variation using HCF. Error free transmission with under 625 ps (under 16 symbol) CDR locking time was achieved without the need to update the transmitted clock phase of incoming optical data packets. The results confirm that HCF provides a robust solution to short-reach optically-switched data centre interconnection in which stable fibre latency enables high-capacity, high throughput, modulation format and data rate agnostic, and low-latency solution. In addition to the benefits from low TCD, HCF offers the advantages of an ultra-wide transmission window, much greater than 100 nm [112], promising potential coarse wavelength division multiplexing (CWDM) for low-cost 800 Gb/s and above data interconnection. Other additional features, including low non-linearity and low chromatic dispersion across a broad wavelength range, offers reduced transmission impairment for future high data rate (50 GBd and beyond) data centre interconnection [120].

7.7 Contribution Statement

The analytical modelling presented in this chapter was conceived, constructed and numerically evaluated by K.A.C., supervised by Z.L.. The experimental work presented in this chapter was originally published as a regular conference paper at ECOC 2019 [109] and as an invited, highly-scored journal paper in JLT [2]. In these papers, the following contributions were made: K.A.C. and Z.L. conceived the concept of single calibration CSA-CDR using low thermal sensitivity fibre, such as HCF. Y.C., E.R.N.F., T.B., F.P., D.J.R. and R.S. designed and manufactured the HC-PBGF used in the experiment. K.A.C. implemented CSA-CDR in the experiment. K.A.C. designed and implemented all FPGA hardware code. K.A.C. led the experiment and collected all experimental results, supervised by Z.L. and R.S.. K.A.C., Z.L. and R.S. wrote and revised the paper manuscript. All authors discussed the results and commented on the paper manuscript. The paper manuscript has been heavily altered for presentation in this thesis.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

Increasing the capacity of data centre networks while minimising their power consumption is key to continuing to support the rapid growth of between server traffic in data centres, and to minimising the impact of data centre power consumption on global warming. All-optical switches, where data is switched in the optical domain rather than in the electronic domain, is a promising potential technology for achieving these goals [7, 8]. However, to enable high-performance all-optical data centre switches, CDR locking time must be minimised to sub-nanosecond.

Chapter 2 focused on exploring the data centre environment, including an introduction to the current electronically-switched data centre architecture, consideration of current standardised methods of transmission, the data centre thermal environment, as well as summarising approaches to implementing optical switching within the data centre. The nanosecond-switching time all-optically switched approach to data centre optical switching avoids the scheduling concerns associated with hybrid approaches, and as long as CDR locking time is minimised to sub-nanosecond to complement recent advances in optical switch reconfiguration times [13, 60], all-optical switches are able to handle real data centre traffic with high performance [8].

Chapter 3 then explored previous approaches to burst-mode CDR. CDRs based on GVCO have been shown to achieve sub-nanosecond CDR locking time, but these CDRs have practical limitations identified in literature that render them inappropriate for data centre optical switching. Recent research has focused on minimising the CDR locking time of digital PI based CDRs, but the state-of-the-art CDR locking time of such CDRs has been limited to only 325 symbols (equivalent to 12.7 ns at 25.6 Gb/s) [12]. This limitation arises from the need for these CDRs to operate with a range of different clock frequency offsets and any clock phase offset. As was shown later in this thesis, this limitation can be circumvented by phase and frequency synchronising all end-points connected to an optical switch, leveraging the requirement for synchronisation of end-points connected to optical switches to avoid packet collisions.

Chapters 4 and *5* established an analytical model of single calibration CSA-CDR without and with packet clock phase tracking respectively. In both of these approaches, all end-points connected to an optical switch are frequency synchronised, and phase synchronisation of all end-points is then established on network startup. In single calibration CSA-CDR without packet clock phase tracking, the CDR is then unused to minimise power consumption, and in single calibration CSA-CDR with packet clock phase tracking, the CDR is still used after phase synchronisation for locking to incoming data packets. The analytical model established in these Chapters was built to explore the impact of data centre temperature variation, as well as to model the performance characteristics of realistic data centre transmitters and receivers. The outcome of the modelling is that for both approaches, if SMF-28 transmission is used, they are only feasible for intra-rack transmission, and that further refinement is required to improve the practicality of the approaches.

Chapter 6 explored the impact of extending single calibration CSA-CDR to regularly updating the clock phase values, in an approach to CSA-CDR called *clock phase caching*. The analytical modelling established in *Chapters 4* and *5* was extended to model the impact of regular clock phase updates on performance and transmission overhead. The outcome of this modelling was that clock phase caching was determined to be potentially feasible. An experimental prototype of clock phase caching then demonstrated that the approach can be used to achieve under 625 ps CDR locking time for optically-switched data-centre networks. The clock phase caching technique was shown to tolerate a rate-of-change of temperature of 0.11 °C/s, which was 3 times greater than the worst-case conditions observed in a production cloud data centre, as well as a root-mean-square clock source jitter tolerance of 0.135 symbols. The technique allows for an increase of data-centre optical switch throughput to more than 90% versus current previous approaches to CDR. The approach could be scaled to support 10,000 data-centre end-points (servers or switches) for realistic worst case data-centre environmental temperature variation and fibre lengths with a small worst-case 1.7% overhead on network throughput.

Chapter 7 explored the impact of using HCF transmission on single calibration CSA-CDR. It was demonstrated that clock-synchronised data center systems using both approaches can have 20 times greater tolerance to temperature variation using HCF versus the SMF-28 transmission approach in *Chapters 4* and *5*. Error free transmission with under 625 ps CDR locking time was achieved without the need to actively track the clock phase of incoming optical packets. These results confirm that HCF provides a robust solution to short-reach optically-switched data centre interconnection, allowing single calibration CSA-CDR to maintain sub-nanosecond CDR locking time within an optically-switched data centre cluster with fibre interconnection lengths of up to 100 m, across a 22.7 °C temperature range.

This thesis proposed, analytically modelled and experimentally demonstrated the clock synchronisation assisted clock and data recovery (CSA-CDR) approach to burst-mode CDR, which uses the measurement and storage of clock phase values atop a clock frequency synchronised network to simplify the process of CDR. The approach allows practical sub-nanosecond CDR locking time to be achieved in a data centre context for the first time, enabling the performance and power consumption benefits of data centre all-optical switches to be realised.

8.2 Future Work

There are many potential future research directions that could build on the work demonstrated in this thesis. This section will explore these future directions. The key areas of focus for further research into CSA-CDR may be summarised as:

- Support of transmission contexts beyond intra-data centre interconnection.
- Increasing transmission distance.
- Support of higher symbol rates.
- Support of higher-order modulation formats.
- Reduction of overhead.
- Investigation of benefits of multi-core fibre.
- Time synchronisation.

These key areas will now be discussed in more detail.

8.2.1 Supporting Longer Distance Scale Applications

This thesis demonstrated the use of CSA-CDR in the context of data centre optical switching, which is a temperature controlled environment with maximum within-building fibre lengths of 2 km. Although single calibration CSA-CDR is not likely to be practical beyond distance scales of 100 m, clock phase caching could potentially be applied at longer length scales, such as in inter-data centre optical networks [32] or time division multiplexed passive optical networks (TDM-PONs) [121]. Inter-data centre optical networks are of interest for creating virtual large scale data centres that serve large metropolitan areas while minimising the capital cost associated with building a single data centre of an equivalent size on a single site [122]. TDM-PONs, as shown in Figure 8.1, are of interest for access applications, including residential and commercial internet connection using fibre to the home (FTTH) and for 5G base station interconnection.

Both of these applications could benefit from the clock phase caching approach demonstrated in this thesis, by enabling sub-nanosecond CDR locking time while potentially enabling time synchronisation to be built atop the clock phase caching technique (which will be discussed in more detail later in this section). Typical standardised fibre lengths in both of these applications are up to 40 km, approximately 20 times longer than the 2 km maximum fibre length for intra-data centre interconnection, requiring at least a 20 times faster clock phase update rate for the same rate-of-change of temperature. The maximum supported number of end-points interconnected by a TDM-PON or inter-data centre building network would be dependent on the worst-case rate-of-change of fibre temperature. TDM-PONs must support 256 end-points [121], which is much smaller than the 10,000 end-points considered for intra-data centre optical switching within this thesis, but the worst-case change in fibre temperature is not known. Future research would need to measure and/or calculate (based on environmental temperature records) the worst-case change in fibre temperature across these long fibre transmission lengths. Future research would also need to investigate the impact of wavelength dispersion on the maximum tolerable clock phase offset. If the worst-case rate-of-change of temperature and dispersion impairments are sufficiently small or easily compensatable, clock phase caching could be a promising approach for CDR in these systems.

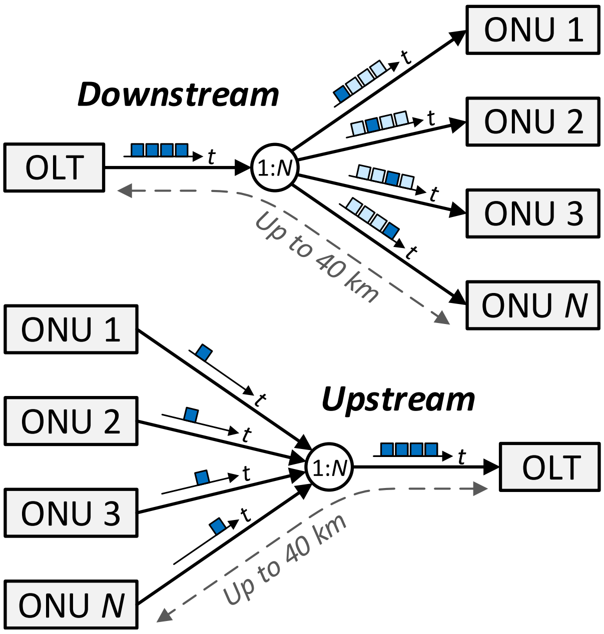

Fig. 8.1: TDM-PON upstream and downstream links, a method of implementing FTTH broadband. OLT, optical line terminal (located within internet provider exchange); ONU, optical network unit (located within residences or businesses). The upstream transmissions are time division multiplexed, and so must be synchronised. Synchronisation inaccuracy of the transmissions, and the time taken to perform CDR locking upon reception of a new transmission, leads to transmission overhead, both problems which could potentially be addressed with CSA-CDR.

8.2.2 Supporting Higher Symbol Rates and Higher Order Modulation Formats

The work presented in this thesis used 25 GBaud NRZ-OOK. The most recent intra-data centre interconnection standards use 25 GBaud PAM-4 [10], and those that follow within the next 5 years are likely to use either SDM or CWDM with 50 GBaud PAM-4 or 25 GBaud PAM-8 to reach over 800 Gb/s per-link transmission rates. The principle of clock phase assisted CDR is applicable to higher order symbol rates and modulation formats as long as clock jitter is minimised, and indeed clock phase caching has recently been demonstrated to operate successfully with 25 GBaud PAM-4 transmission with a maximum sampling phase deviation of ±5 ps (0.125 symbols at 25 GBaud) over 24 hours of measurement [8]. A cyclic scheduler was used in Ballani et al. to avoid the concern of transmission overhead associated with other approaches to scheduling [8]. However, the overall number of supportable end-points has not been quantified, and the extension of clock phase caching to symbol rates greater than 25 GBaud has not yet been analytically or experimentally investigated.

Future research should investigate these transmission scenarios using a combination of analytical and experimental work. Minimising jitter between the sampling clock and incoming data, which was a key factor limiting performance in this thesis, will be critical for maximising performance with higher order modulation formats such as PAM-4 and higher symbol rates. Future research should also investigate methods of minimising the required minimum clock phase update rate for a given rate-of-change of temperature, which in turn would minimise transmission overhead. Potential avenues for investigation could include: extending the clock phase caching algorithm to include consideration of the long-term trend in clock phase values (essentially adding a second order clock phase change loop), allowing the clock phase to be reset by a clock phase update to a clock phase offset value other than the central sampling point at ϕ = 0 symbols, using machine learning to predict required changes in clock phase values based on temperature changes within the data centre, and further research to build on existing research into performing clock phase synchronisation with low thermal sensitivity fibres such as HCF [2] and MCF [123].

8.2.3 Time Synchronisation

An additional application of the clock phase caching approach to CSA-CDR in optical switches is time synchronisation. The distributed clock phase synchronisation that is key to the operation of CSA-CDR operates by tracking all fibre delay changes that occur through optical fibres that interconnect nodes connected through an optical switch. Once initial clock phase synchronisation is established, all interconnected nodes maintain relative clock phase synchronisation with each other. This can be leveraged to perform time synchronisation, which has been demonstrated as part of the Sirius optical switch demonstrator shown in Ballani et al., which used the clock phase caching approach to CSA-CDR [8].

CSA-CDR could, therefore, be of value in the various scientific and engineering applications that require synchronisation. For example, TDM-PONs could potentially benefit from the synchronisation technology demonstrated in this thesis [121]. The optical network units in TDM-PONs can be synchronised in a similar manner to minimise the required inter-packet gap and data-detection latency, which is crucial for latency-sensitive applications such as the Internet of Skills, virtual reality and gaming [124]. Another example lies in clock synchronisation for quantum key distribution (QKD) systems, which require the time gate to be synchronised to the photon arrival, to identify the quantum signals correctly [125]. The clock phase caching approach to CSA-CDR could track the drift of the time-of-flight in transmission links without the need for paired fibre, significantly reducing the complexity of QKD clock networks.

Bibliography

[1] Clark, K. A., Cletheroe, D., Gerard, T., Haller, I., Jozwik, K., Shi, K., Thomsen, B., Williams, H., Zervas, G., Ballani, H., Bayvel, P., Costa, P. and Liu, Z. Synchronous subnanosecond clock and data recovery for optically switched data centres using clock phase caching. *Nature Electronics* **3**, 1–13 (2020).

[2] Clark, K. A., Chen, Y., Fokua, E. R. N., Bradley, T., Poletti, F., Richardson, D. J., Bayvel, P., Slavík, R. and Liu, Z. Low Thermal Sensitivity Hollow Core Fiber for Optically-Switched Data Centers. *Journal of Lightwave Technology* **38**, 2703–2709 (2020).

[3] Cisco. Cisco Annual Internet Report, 2018-2023 (2020). URL `https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf`. Accessed: 14-01-2021.

[4] Singh, A., Ong, J., Agarwal, A., Anderson, G., Armistead, A., Bannon, R., Boving, S., Desai, G., Felderman, B., Germano, P. *et al.* Jupiter rising: A decade of clos topologies and centralized control in Google's datacenter network. *ACM SIGCOMM computer communication review* **45**, 183–197 (2015).

[5] Zilberman, N., Bracha, G. and Schzukin, G. Stardust: Divide and conquer in the data center network. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19)*, 141–160 (2019).

[6] Markov, I. L. Limits on fundamental limits to computation. *Nature* **512**, 147 (2014).

[7] Ballani, H., Costa, P., Haller, I., Jozwik, K., Shi, K., Thomsen, B. and Williams, H. Bridging the Last Mile for Optical Switching in Data Centers. In *Optical Fiber Communication Conference (OFC'18)* (San Diego, CA, USA, 2018).

[8] Ballani, H., Costa, P., Behrendt, R., Cletheroe, D., Haller, I., Jozwik, K., Karinou, F., Lange, S., Shi, K., Thomsen, B. and Williams, H. Sirius: A Flat Datacenter Network with Nanosecond Optical Switching. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication*

on the applications, technologies, architectures, and protocols for computer communication, 782–797 (ACM, New York City, NY, USA, 2020).

[9] Ballani, H., Costa, P., Behrendt, R., Cletheroe, D., Haller, I., Jozwik, K., Karinou, F., Lange, S., Shi, K., Thomsen, B. and Williams, H. Sirius: A Flat Datacenter Network with Nanosecond Optical Switching. SIGCOMM 2020 full talk (2020). URL `https://www.microsoft.com/en-us/research/video/sirius-a-flat-datacenter-network-with-nanosecond-optical-switching-2`. Accessed: 06-01-2021.

[10] IEEE. 802.3-2018 IEEE Standard for Ethernet 802.3-2018 (2018). URL `https://standards.ieee.org/standard/802_3-2018.html`. Accessed: 14-01-2021.

[11] Andrae, A. and Edler, T. On Global Electricity Usage of Communication Technology: Trends to 2030. *Challenges* **6**, 117–157 (2015).

[12] Ozkaya, I., Cevrero, A., Francese, P. A., Menolfi, C., Braendli, M., Morf, T., Kuchta, D., Kull, L., Kossel, M., Luu, D., Meghelli, M., Leblebici, Y. and Toifl, T. A 56Gb/s burst-mode NRZ optical receiver with 6.8ns power-on and CDR-Lock time for adaptive optical links in 14nm FinFET CMOS. *Digest of Technical Papers - IEEE International Solid-State Circuits Conference* **61**, 266–268 (2018).

[13] Shi, K., Lange, S., Haller, I., Cletheroe, D., Behrendt, R., Thomsen, B., Karinou, F., Jozwik, K., Costa, P. and Ballani, H. System Demonstration of Nanosecond Wavelength Switching with Burst-mode PAM4 Transceiver. In *2019 European Conference on Optical Communication (ECOC)*, 1–3 (IEEE, Dublin, Ireland, 2019).

[14] Cheng, Q., Wonfor, A., Wei, J. L., Penty, R. V. and White, I. H. Low-energy, high-performance lossless 8x8 SOA switch. In *OSA/IEEE OFC* (2015).

[15] Chen, C. P., Zhu, X., Liu, Y., Wen, K., Chik, M. S., Baehr-Jones, T., Hochberg, M. and Bergman, K. Programmable Dynamically-Controlled Silicon Photonic Switch Fabric. *Journal of Lightwave Technology* **34** (2016).

[16] Serrano, J. The White Rabbit Project. In *IBIC* (Oxford, UK, 2013).

[17] Roy, A., Zeng, H., Bagga, J., Porter, G. and Snoeren, A. C. Inside the Social Network's (Datacenter) Network. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, SIGCOMM '15, 123–137 (Association for Computing Machinery, New York City, NY, USA, 2015).

[18] IEEE. 802.3az-2010 (Amendment to IEEE Std 802.3-2008) IEEE Standard for Information technology– Local and metropolitan area networks– Specific requirements– Part 3: CSMA/CD Access Method and Physical Layer Specifications Amendment 5: Media Access Control Parameters, Physical Layers, and Management Parameters for Energy-Efficient Ethernet (2010). URL `https://standards.ieee.org/standard/802_3az-2010.html`. Accessed: 14-01-2021.

[19] Clark, K., Ballani, H., Bayvel, P., Cletheroe, D., Gerard, T., Haller, I., Jozwik, K., Shi, K., Thomsen, B., Watts, P., Williams, H., Zervas, G., Costa, P. and Liu, Z. Sub-nanosecond clock and data recovery in an optically-switched data centre network. In *2018 European Conference on Optical Communication (ECOC)*, 1–3 (IEEE, Rome, Italy, 2018).

[20] Zhang, Q., Liu, V., Zeng, H. and Krishnamurthy, A. High-resolution Measurement of Data Center Microbursts. In *Internet Measurement Conference (IMC)* (London, UK, 2017).

[21] Slavík, R., Marra, G., Fokoua, E. N., Baddela, N., Wheeler, N. V., Petrovich, M., Poletti, F. and Richardson, D. J. Ultralow thermal sensitivity of phase and propagation delay in hollow core optical fibres. *Scientific reports* **5**, 15447 (2015).

[22] ASHRAE TC9.9. Data Center Power Equipment Thermal Guidelines and Best Practices (2016). URL `http://tc0909.ashraetcs.org/documents/ASHRAE_TC0909_Power_White_Paper_22_June_2016_REVISED.pdf`. Accessed: 14-01-2021.

[23] Broadcomm. Bcm56990 series. URL `https://www.broadcom.com/products/ethernet-connectivity/switching/strataxgs/bcm56990-series`. Accessed: 28-10-2020.

[24] Al-Fares, M., Loukissas, A. and Vahdat, A. A scalable, commodity data center network architecture. *ACM SIGCOMM Computer Communication Review* **38**, 63 (2008).

[25] Dally, W. J. and Towles, B. P. *Principles and Practices of Interconnection Networks* (Morgan Kaufmann, San Francisco, CA, USA, 2003).

[26] Guo, C., Chen, H., Lin, Z.-W., Kurien, V., Yuan, L., Xiang, D., Dang, Y., Huang, R., Maltz, D., Liu, Z., Wang, V. and Pang, B. Pingmesh: A Large-Scale System for Data Center Network Latency Measurement and Analysis. In *Proceedings of*

the 2015 ACM Conference on Special Interest Group on Data Communication - SIGCOMM 2015, vol. 45, 139–152 (ACM Press, London, UK, 2015).

[27] Nvidia. NVIDIA A100 Tensor Core GPU Architecture (2020). URL `https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf`. Accessed: 28-10-2020.

[28] Taylor, M. B. Is Dark Silicon Useful? Harnessing the Four Horsemen of the Coming Dark Silicon Apocalypse. In *Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE*, 1131–1136 (IEEE, San Francisco, CA, USA, 2012).

[29] Cheng, Q., Bahadori, M., Glick, M., Rumley, S. and Bergman, K. Recent advances in optical technologies for data centers: a review. *Optica* **5**, 1354 (2018).

[30] IEEE. 802.3an-2006 IEEE Standard for Information Technology - Telecommunications and Information Exchange Between Systems âĂŞ LAN/MAN - Specific Requirements Part 3: CSMA/CD Access Method and Physical Layer Specifications - Amendment: Physical Layer and Management Parameters (2006). URL `https://standards.ieee.org/standard/802_3an-2006.html`. Accessed: 14-01-2021.

[31] IEEE. 802.3ba-2010 IEEE Standard for Information technology - Local and metropolitan area networks - Specific requirements - Part 3: CSMA/CD Access Method and Physical Layer Specifications Amendment 4: Media Access Control Parameters, Physical Layers, and Management Parameters for 40 Gb/s and 100 Gb/s Operation (2010). URL `https://standards.ieee.org/standard/802_3ba-2010.html`. Accessed: 14-01-2021.

[32] IEEE. 802.3cn-2019 IEEE Standard for Ethernet - Amendment 4 - Physical Layers and Management Parameters for 50Gb/s, 200Gb/s, and 400Gb/s Operation over Single-Mode Fiber (2019). URL `https://standards.ieee.org/standard/802_3cn-2019.html`. Accessed: 14-01-2021.

[33] IEEE. 802.3cm-2020 IEEE Standard for Ethernet - Amendment 7 - Physical Layer and Management Parameters for 400 Gb/s over Multimode Fiber (2020). URL `https://standards.ieee.org/standard/802_3cm-2020.html`. Accessed: 14-01-2021.

[34] Petrilla, J., Cole, C., King, J., Lewis, D., Hiramoto, K. and Tsumura, E. 100G CWDM4 MSA Specifications Rev 1.1 (2015). URL

https://www.color-chip.com/wp-content/uploads/2018/02/CWDM4-MSA-Technical-Spec-1p1-1.pdf. Accessed: 14-01-2021.

[35] IEEE. 802.3bs-2017 IEEE Standard for Ethernet - Amendment 10 - Media Access Control Parameters, Physical Layers, and Management Parameters for 200 Gb/s and 400 Gb/s Operation (2017). URL https://standards.ieee.org/standard/802_3bs-2017.html. Accessed: 14-01-2021.

[36] IEEE. 802.3bm-2015 IEEE Standard for Ethernet - Amendment 3 - Physical Layer Specifications and Management Parameters for 40 Gb/s and 100 Gb/s Operation over Fiber Optic Cables (2015). URL https://standards.ieee.org/standard/802_3bm-2015.html. Accessed: 14-01-2021.

[37] Kuchta, D. M., Rylyakov, A. V., Doany, F. E., Schow, C. L., Proesel, J. E., Baks, C. W., Westbergh, P., Gustavsson, J. S. and Larsson, A. A 71-Gb/s NRZ Modulated 850-nm VCSEL-Based Optical Link. *IEEE Photonics Technology Letters* **27**, 577–580 (2015).

[38] Lavrencik, J., Varughese, S., Gustavsson, J. S., Haglund, E., Larsson, A. and Ralph, S. E. Error-Free 100Gbps PAM-4 Transmission over 100m Wideband Fiber using 850nm VCSELs. In *2017 European Conference on Optical Communication (ECOC)*, 1–3 (IEEE, Gothenburg, Sweden, 2017).

[39] Kottke, C., Caspar, C., Jungnickel, V., Freund, R., Agustin, M. and Ledentsov, N. N. High Speed 160 Gb/s DMT VCSEL Transmission Using Pre-equalization. In *Optical Fiber Communication Conference*, W4I.7 (OSA, Washington, DC, USA, 2017).

[40] Lavrencik, J., Varughese, S., Thomas, V. A., Landry, G., Sun, Y., Shubochkin, R., Balemarthy, K., Tatum, J. and Ralph, S. E. $4\lambda\times$100Gbps VCSEL PAM-4 Transmission over 105m of Wide Band Multimode Fiber. In *Optical Fiber Communication Conference*, Tu2B.6 (OSA, Washington, DC, USA, 2017).

[41] Pang, X., Ozolins, O., Lin, R., Zhang, L., Udalcovs, A., Xue, L., Schatz, R., Westergren, U., Xiao, S., Hu, W., Jacobsen, G., Popov, S. and Chen, J. 200 Gbps/Lane IM/DD Technologies for Short Reach Optical Interconnects. *Journal of Lightwave Technology* **38**, 492–503 (2020).

[42] Elbers, J.-P., Eiselt, N., Dochhan, A., Rafique, D. and Grießer, H. PAM4 vs Coherent for DCI Applications. In *Advanced Photonics 2017 (IPR, NOMA, Sensors, Networks, SPPCom, PS)*, SpTh2D.1 (Optical Society of America, 2017).

[43] Chen, X., Milosevic, M. M., Stanković, S., Reynolds, S., Bucio, T. D., Li, K., Thomson, D. J., Gardes, F. and Reed, G. T. The Emergence of Silicon Photonics as a Flexible Technology Platform. *Proceedings of the IEEE* **106**, 2101–2116 (2018).

[44] Xilinx. UG578 (v1.3). UltraScale Architecture GTY Transceivers - User Guide (2017). URL `https://www.xilinx.com/support/documentation/user_guides/ug578-ultrascale-gty-transceivers.pdf`. Accessed: 14-01-2021.

[45] Kachris, C. and Tomkos, I. A Survey on Optical Interconnects for Data Centers. *IEEE Communications Surveys and Tutorials* **14**, 1021–1036 (2012).

[46] Andreades, P. *Control Plane Hardware Design for Optical Packet Switched Data Centre Networks*. Ph.D. thesis, UCL (2019).

[47] Benjamin, J. L. *Towards Sub-microsecond Optical Circuit Switched Networks for Future Data Centers*. Ph.D. thesis, UCL (2020).

[48] Wang, G., Andersen, D. G., Kaminsky, M., Papagiannaki, K., Ng, T. E., Kozuch, M. and Ryan, M. c-Through: Part-time Optics in Data Centers. *ACM SIGCOMM Computer Communication Review* **40**, 327–338 (2010).

[49] Farrington, N., Porter, G., Radhakrishnan, S., Bazzaz, H. H., Subramanya, V., Fainman, Y., Papen, G. and Vahdat, A. Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers. In *Proceedings of the ACM SIGCOMM 2010 conference on SIGCOMM - SIGCOMM 2010*, 339 (ACM Press, New York City, NY, USA, 2010).

[50] Porter, G., Strong, R., Farrington, N., Forencich, A., Chen-Sun, P., Rosing, T., Fainman, Y., Papen, G. and Vahdat, A. Integrating microsecond circuit switching into the data center. In *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM - SIGCOMM '13*, 447 (ACM Press, New York City, NY, USA, 2013).

[51] Polatis. Series 7000n (2017). URL `https://www.polatis.com/series-7000-384x384-port-software-controlled-optical-circuit-switch-sdn-enabled.asp`. Accessed: 28-10-2020.

[52] Bazzaz, H. H., Tewari, M., Wang, G., Porter, G., Ng, T. S. E., Andersen, D. G., Kaminsky, M., Kozuch, M. A. and Vahdat, A. Switching the Optical Divide: Fundamental Challenges for Hybrid Electrical/Optical Datacenter Networks. In *Proceedings of the 2nd ACM Symposium on Cloud Computing*, 1–8 (ACM, Cascais, Portugal, 2011).

[53] Liu, H., Lu, F., Forencich, A., Kapoor, R., Tewari, M., Voelker, G. M., Papen, G., Snoeren, A. C. and Porter, G. Circuit switching under the radar with reactor. In *Proceedings of the 11th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2014*, 1–15 (USENIX, Seattle, WA, USA, 2014).

[54] Mellette, W. M., Snoeren, A. C. and Porter, G. Toward Optical Switching in the Data Center (Invited Paper). In *2018 IEEE 19th International Conference on High Performance Switching and Routing (HPSR)*, vol. 2018-June, 1–6 (IEEE, Bucharest, Romania, 2018).

[55] Shacham, A., Small, B., Liboiron-Ladouceur, O. and Bergman, K. A fully implemented 12×12 data vortex optical packet switching interconnection network. *Journal of Lightwave Technology* **23**, 3066–3075 (2005).

[56] Shacham, A. and Bergman, K. An Experimental Validation of a Wavelength-Striped, Packet Switched, Optical Interconnection Network. *Journal of Lightwave Technology* **27**, 841–850 (2009).

[57] Proietti, R., Nitta, C. J., Akella, V. and Yoo, S. J. B. LIONS: An AWGR-Based Low-Latency Optical Switch for High-Performance Computing and Data Centers. *IEEE Journal of Selected Topics in Quantum Electronics* **19**, 3600409–3600409 (2013).

[58] Alexoudi, T., Kanellos, G. T. and Pleros, N. Optical RAM and integrated optical memories: a survey. *Light: Science and Applications* **9**, 91 (2020).

[59] Hemenway, R. and Grzybowski, R. R. An Optical Packet-Switched Interconnect for Supercomputer Applications. *Journal of Optical Networking* **3**, 900–913 (2004).

[60] Gerard, T., Dzieciol, H., Benjamin, J., Clark, K., Williams, H., Thomsen, B., Lavery, D. and Bayvel, P. Packet Timescale Wavelength Switching Enabled by Regression Optimisation. *IEEE Photonics Technology Letters* **32**, 477–480 (2020).

[61] Bowers, J. E., Komljenovic, T., Davenport, M., Hulme, J., Liu, A. Y., Santis, C. T., Spott, A., Srinivasan, S., Stanton, E. J. and Zhang, C. Recent advances in silicon photonic integrated circuits. In Li, G. and Zhou, X. (eds.) *Next-Generation Optical Communication: Components, Sub-Systems, and Systems V*, vol. 9774, 977402 (2016).

[62] Benjamin, J. L., Gerard, T., Lavery, D., Bayvel, P. and Zervas, G. PULSE: Optical Circuit Switched Data Center Architecture Operating at Nanosecond Timescales. *Journal of Lightwave Technology* **38**, 4906–4921 (2020).

[63] Andreades, P. and Zervas, G. Parallel Modular Scheduler Design for Clos Switches in Optical Data Center Networks. *Journal of Lightwave Technology* **38**, 3506–3518 (2020).

[64] Shrivastav, V., Valadarsky, A., Ballani, H., Costa, P., Lee, K. S., Wang, H., Agarwal, R. and Weatherspoon, H. Shoal: A network architecture for disaggregated racks. In *Proceedings of the 16th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2019*, 255–270 (USENIX, Boston, MA, USA, 2019).

[65] Bostica, B., Burzio, M., Gambini, P. and Zucchelli, L. Synchronisation issues in optical packet switched networks. In Prati, G. (ed.) *Photonic Networks*, 362–376 (Springer London, London, 1997).

[66] Wire Raven Blog. Approaches to hot and cold air aisle containment and what to know (2017). URL `http://wireraven.com/hot-vs-cold-aisle-containment-what-you-need-know/`. Accessed: 28-10-2020.

[67] ANSI/TIA. ANSI/TIA-942-B Telecommunications Infrastructure Standard for Data Centers (2017). URL `https://global.ihs.com/doc_detail.cfm?document_name=TIA%2D942&item_s_key=00414811`. Accessed: 14-01-2020.

[68] ASHRAE TC9.9. Data Center Networking Equipment–Issues and Best Practices (2013). URL `https://tc0909.ashraetcs.org/documents/ASHRAE%20Networking%20Thermal%20Guidelines.pdf`. Accessed: 14-01-2021.

[69] Rylyakov, A., Proesel, J. E., Rylov, S., Lee, B. G., Bulzacchelli, J. F., Ardey, A., Parker, B., Beakes, M., Baks, C. W., Schow, C. L. and Meghelli, M. A 25 Gb/s Burst-Mode Receiver for Low Latency Photonic Switch Networks. *IEEE Journal of Solid-State Circuits* **50**, 3120–3132 (2015).

[70] Banu, M. and Dunlop, A. E. Clock recovery circuits with instantaneous locking. *Electronics Letters* **28**, 2127–2130 (1992).

[71] Hsieh, M.-t. and Sobelman, G. Architectures for multi-gigabit wire-linked clock and data recovery. *IEEE Circuits and Systems Magazine* **8**, 45–57 (2008).

[72] Cho, L.-C., Lee, C. and Liu, S.-I. A 33.6-to-33.8Gb/s Burst-Mode CDR in 90nm CMOS. In *2007 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, 48–586 (IEEE, San Francisco, CA, USA, 2007).

[73] Terada, J., Nishimura, K., Kimura, S., Katsurai, H., Yoshimoto, N. and Ohtomo, Y. A 10.3 Gb/s burst-mode CDR Using a $\Delta\Sigma$ DAC. *IEEE Journal of Solid-State Circuits* **43**, 2921–2928 (2008).

[74] Nogawa, M., Nishimura, K., Kimura, S., Yoshida, T., Kawamura, T., Togashi, M., Kumozaki, K. and Ohtomo, Y. A 10Gb/s burst-mode CDR IC in 0.13μm CMOS. *Digest of Technical Papers - IEEE International Solid-State Circuits Conference* **48**, 228–595 Vol. 1 (2005).

[75] Lee, J. and Liu, M. A 20-Gb/s Burst-Mode Clock and Data Recovery Circuit Using Injection-Locking Technique. *IEEE Journal of Solid-State Circuits* **43**, 619–630 (2008).

[76] Shastri, B. J. and Plant, D. V. 5/10-Gb/s burst-mode clock and data recovery based on semiblind oversampling for PONs: Theoretical and experimental. *IEEE Journal on Selected Topics in Quantum Electronics* **16**, 1298–1320 (2010).

[77] Shastri, B. J. and Plant, D. V. A 10-Gb/s space sampling burst-mode clock and data recovery circuit for passive optical networks. *IEEE Photonic Society 24th Annual Meeting, PHO 2011* **3**, 937–938 (2011).

[78] Xilinx. DS893 (v1.12). Virtex UltraScale Architecture Data Sheet: DC and AC Switching Characteristics (2019). URL `https://www.xilinx.com/support/documentation/data_sheets/ds893-virtex-ultrascale-data-sheet.pdf`. Accessed: 14-01-2021.

[79] Blank, L. C., Bryant, E. G., Lord, A., Boggis, J. M. and Stallard, W. A. 150km Optical Fibre Transmission Network Experiment with 2Gbit/s Throughput. *Electronics Letters* **23**, 977–978 (1987).

[80] Lord, A., Blank, L. C., Boggis, J. M., Bryant, E. and Stallard, W. A. Optical multiplexing techniques for future Gbit/s transmission systems. In *IEEE International Conference on Communications, - Spanning the Universe.*, 21–25 (IEEE, Philadelphia, PA, USA, 1988).

[81] Lord, A., Blank, L. C., Boggis, J. M., Bryant, E. and Stallard, W. A. Theory of Control Mechanism for an Optically Time-Division-Multiplexed System. *Electronics Letters* **24**, 2011–2012 (1988).

[82] Ellis, A. D., Widdowson, T., Phillips, I. D. and Pender, W. A. High speed OTDM networks employing electro-optic modulators. *IEICE Transactions on Electronics* **E81-C**, 1301–1308 (1998).

[83] ITU-T. G.8261 Timing and synchronization aspects in packet networks (2008). URL `https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-G.8261-201908-I!!PDF-E&type=items`. Accessed: 14-01-2021.

[84] Ma, L.-S., Jungner, P., Ye, J. and Hall, J. L. Delivering the same optical frequency at two places: accurate cancellation of phase noise introduced by an optical fiber or other time-varying path. *Optics Letters* **19**, 1777 (1994).

[85] Lopez, O., Kéfélian, F., Jiang, H., Haboucha, A., Bercy, A., Stefani, F., Chanteau, B., Kanj, A., Rovera, D., Achkar, J., Chardonnet, C., Pottie, P.-E., Amy-Klein, A. and Santarelli, G. Frequency and time transfer for metrology and beyond using telecommunication network fibres. *Comptes Rendus Physique* **16**, 531–539 (2015).

[86] Agrawal, G. P. *Fiber-Optic Communication Systems* (John Wiley and Sons, New Jersey, USA, 2010), 4th edn.

[87] ANSI. T1.105.07-1996 (R2005) Synchronous Optical Network (SONET) âĂŞ Sub-STS-1 Interface Rates and Formats Specification (2005). URL `https://webstore.ansi.org/standards/atis/ansit1105071996r2005`. Accessed: 14-01-2021.

[88] ITU-T. G.707 Network node interface for the synchronous digital hierarchy (SDH) (2007). URL `http://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-G.707-200701-I!!PDF-E&type=items`. Accessed: 14-01-2021.

[89] IEEE. 1588-2019 IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems (2019). URL `https://standards.ieee.org/standard/1588-2019.html`. Accessed: 14-01-2020.

[90] Lee, K. S., Wang, H., Shrivastav, V. and Weatherspoon, H. Globally Synchronized Time via Datacenter Networks. In *ACM SIGCOMM Conference on Data Communication* (Florianópolis, Brazil, 2016).

[91] Aliaga, R. J., Monzo, J. M., Spaggiari, M., Ferrando, N., Gadea, R. and Colom, R. J. PET system synchronization and timing resolution using high-speed data links. *IEEE Transactions on Nuclear Science* **58**, 1–7 (2011).

[92] Lee, K. S., Wang, H., Shrivastav, V. and Weatherspoon, H. Globally synchronized time via datacenter networks. In *Proceedings of the 2016 ACM SIGCOMM Conference*, 454–467 (ACM, Florianópolis, Brazil, 2016).

[93] Soleimani, S., Afzali-Kusha, A. and Forouzandeh, B. Temperature dependence of propagation delay characteristic in FinFET circuits. *Proceedings of the International Conference on Microelectronics, ICM* 276–279 (2008).

[94] Bousonville, M., Bock, M., Felber, M., Ladwig, T., Lamb, T., Schlarb, H., Schulz, S., Sydlo, C., Hunziker, S., Kownacki, P. and Jablonski, S. New phase stable optical fiber. In *Beam Instrumentation Workshop*, 101–103 (Newport News, VA, USA, 2012).

[95] Bottacchi, S. *Noise and Signal Interference in Optical Fiber Transmission Systems: An Optimum Design Approach* (John Wiley and Sons, Chichester, 2008).

[96] Zverev and Blinchikoff. *Filtering in the Time and Frequency Domains* (SciTech Publishing Inc, Raleigh, USA, 2001).

[97] Jeffrey, A. *Handbook of Mathematical Formulas and Integrals* (Academic Press, Inc., San Diego, CA, USA, 1995).

[98] Texas Instruments. ONET2804TLP Low-Power, 28-Gbps, 4-Channel Limiting TIA (2017). URL `https://www.ti.com/lit/ds/sbas796/sbas796.pdf?ts=1595588247747`. Accessed: 28-10-2020.

[99] Beijing Lightsending Technologies Ltd. 25G High Speed InGaAs PIN photodiode (2017). URL `http://www.lightsensing.com/upfile/2017-06/20170622853.pdf`. Accessed: 28-10-2020.

[100] NIST. CODATA Value: Boltzmann Constant (2018). URL `https://physics.nist.gov/cgi-bin/cuu/Value?k`. Accessed: 30-12-2020.

[101] NIST. CODATA Value: Elementary Charge (2018). URL `https://physics.nist.gov/cgi-bin/cuu/Value?e`. Accessed: 30-12-2020.

[102] Zhang, H., Ou, J. and Krooswyk, S. *High Speed Digital Design* (Morgan Kaufmann, Burlington, MA, USA, 2015), 1st edn.

[103] Lipínski, M., Włostowski, T., Serrano, J. and Alvarez, P. White rabbit: a PTP application for robust sub-nanosecond synchronization. In *IEEE International Symposium on Precision Clock Synchronization for Measurement, Control and Communication* (2011).

[104] Torres-Company, V. and Weiner, A. M. Optical frequency comb technology for ultra-broadband radio-frequency photonics. *Laser and Photonics Reviews* **8**, 368–393 (2014).

[105] Kuo, B. P.-P., Myslivets, E., Alic, N. and Radic, S. Wavelength multicasting via frequency comb generation in a bandwidth-enhanced fiber optical parametric mixer. *Journal of Lightwave Technology* **29**, 3515–3522 (2011).

[106] Zhang, M., Buscaino, B., Wang, C., Shams-Ansari, A., Reimer, C., Zhu, R., Kahn, J. M. and Lončar, M. Broadband electro-optic frequency comb generation in a lithium niobate microring resonator. *Nature* **568**, 373 (2019).

[107] van Aardenne-Ehrenfest, T. and De Bruijn, N. *Circuits and Trees in Oriented Linear Graphs* (Birkhäuser, Boston, MA, USA, 2009).

[108] ITU-T. O.150 Digital test patterns for performance measurements on digital transmission equipment (1992). URL `https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-O.150-199210-S!!PDF-E&type=items`. Accessed: 22-12-2020.

[109] Clark, K. A., Chen, Y., Fokua, E. R. N., Bradley, T., Poletti, F., Richardson, D. J., Bayvel, P., Slavík, R. and Liu, Z. Low thermal sensitivity hollow core fiber for optically-switched data centers. In *2019 European Conference on Optical Communication (ECOC)*, 1–3 (IEEE, Dublin, Ireland, 2019).

[110] Fokoua, E. N., Petrovich, M. N., Bradley, T., Poletti, F., Richardson, D. J. and Slavík, R. How to make the propagation time through an optical fiber fully insensitive to temperature variations. *Optica* **4**, 659–668 (2017).

[111] Bradley, T., Hayes, J., Hooper, L., Sakr, H., Jasion, G., Alonso, M., Taranta, A., Saljoghei, A., Fake, M., Davidson, I., Chen, Y., Wheeler, N., Fokoua, E., Wang, W., Sandoghchi, S. R., Richardson, D., Poletti, F. and Mulvad, H. C. Antiresonant hollow core fibre with 0.65 dB/km attenuation across the C and L telecommunication bands. In *2019 European Conference on Optical Communication (ECOC)*, 1–3 (IEEE, Dublin, Ireland, 2019).

[112] Nespola, A., Straullu, S., Bradley, T., Mulvad, H. C., Hayes, J., Jasion, G., Gouveia, M., Sandoghchi, S. R., Bawn, S., Forghieri, F., Richardson, D., Poletti, F. and Poggiolini, P. Record PM-16QAM and PM-QPSK Transmission Distance (125 and 340 km) over Hollow-Core-Fiber. In *2019 European Conference on Optical Communication (ECOC)*, 1–3 (IEEE, Dublin, Ireland, 2019).

[113] Poletti, F., Wheeler, N., Petrovich, M., Baddela, N., Fokoua, E. N., Hayes, J., Gray, D., Li, Z., Slavík, R. and Richardson, D. Towards high-capacity fibre-optic communications at the speed of light in vacuum. *Nature Photonics* **7**, 279 (2013).

[114] Poletti, F., Petrovich, M. N. and Richardson, D. J. Hollow-core photonic bandgap fibers: technology and applications. *Nanophotonics* **2**, 315–340 (2013).

[115] Silicon Laboratories. Si5324 Any-Frequency Precision Clock Multiplier / Jitter Attenuator. URL `https://www.silabs.com/documents/public/data-sheets/Si5324.pdf`. Accessed: 28-10-2020.

[116] Chen, Y., Liu, Z., Sandoghchi, S. R., Jasion, G. T., Bradley, T. D., Numkam Fokoua, E., Hayes, J. R., Wheeler, N. V., Gray, D. R., Mangan, B. J., Slavík, R., Poletti, F., Petrovich, M. N. and Richardson, D. J. Multi-kilometer long, longitudinally uniform hollow core photonic bandgap fibers for broadband low latency data transmission. *Journal of Lightwave Technology* **34**, 104–113 (2016).

[117] Slavík, R., Petrovich, M., Wheeler, N., Hayes, J., Baddela, N., Gray, D., Poletti, F. and Richardson, D. 1.45 Tbit/s, Low Latency Data Transmission through a 19-Cell Hollow Core Photonic Band Gap Fibre. In *2012 European Conference and Exhibition on Optical Communication (ECOC)*, Mo.2.F.2 (OSA, Washington, DC, USA, 2012).

[118] Komanec, M., Suslov, D., Zvánovec, S., Chen, Y., Bradley, T., Sandoghchi, S. R., Fokoua, E. R., Jasion, G. T., Petrovich, M. N., Poletti, F., Richardson, D. J. and Slavík, R. Low-Loss and Low-Back-Reflection Hollow-Core to Standard Fiber Interconnection. *IEEE Photonics Technology Letters* **31**, 723–726 (2019).

[119] Nespola, A., Straullu, S., Bradley, T., Mulvad, H., Hayes, J., Jasion, G., Gouveia, M., Sandoghchi, S., Bawn, S., Forghieri, F., Richardson, D., Poletti, F. and Poggiolini, P. Record PM-16QAM and PM-QPSK transmission distance (125 and 340 km) over hollow-core-fiber. In *2019 European Conference on Optical Communication (ECOC 2019)*, 1–4 (IEEE, Dublin, Ireland, 2019).

[120] Liu, Z., Karanov, B., Galdino, L., Hayes, J. R., Lavery, D., Clark, K., Shi, K., Elson, D. J., Thomsen, B. C., Petrovich, M. N., Richardson, D. J., Poletti, F., Slavík, R. and Bayvel, P. Nonlinearity-free coherent transmission in hollow-core antiresonant fiber. *Journal of Lightwave Technology* **37**, 909–916 (2019).

[121] ITU-T. G.989.2 40-Gigabit-capable passive optical networks 2 (NG-PON2) (2015). URL `https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-G.989.3-201510-I!!PDF-E&type=items`. Accessed: 14-01-2021.

[122] Dukic, V., Khanna, G., Gkantsidis, C., Karagiannis, T., Parmigiani, F., Singla, A., Filer, M., Cox, J. L., Ptasznik, A., Harland, N., Saunders, W. and Belady, C. Beyond the mega-data center. In *Proceedings of the Annual conference of*

the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication, 765–781 (ACM, New York City, NY, USA, 2020).

[123] Sohanpal, R. S., Clark, K. A., Puttnam, B. J., Awaji, Y., Wada, N., Bayvel, P. and Liu, Z. Clock and Data Recovery-Free Data Communications Enabled by Multi-Core Fiber With Low Thermal Sensitivity of Skew. *Journal of Lightwave Technology* **38**, 1636–1643 (2020).

[124] Aijaz, A., Dohler, M., Aghvami, A. H., Friderikos, V. and Frodigh, M. Realizing the tactile Internet: Haptic communications over next generation 5G cellular networks. *IEEE Wireless Communications* **24**, 82–89 (2016).

[125] Takesue, H., Nam, S. W., Zhang, Q., Hadfield, R. H., Honjo, T., Tamaki, K. and Yamamoto, Y. Quantum key distribution over a 40-dB channel loss using superconducting single-photon detectors. *Nature Photonics* **1**, 343 (2007).

[126] Jri Lee, Kundert, K. and Razavi, B. Analysis and modeling of bang-bang clock and data recovery circuits. *IEEE Journal of Solid-State Circuits* **39**, 1571–1580 (2004).

[127] Widmer, A. X. and Franaszek, P. A. A DC-Balanced, Partitioned-Block, 8B/10B Transmission Code. *IBM Journal of Research and Development* **27**, 440–451 (1983).

Appendices

Appendix A

Mathematical Derivations

This appendix provides mathematical derivations for several of the equations given earlier in the thesis. These derivations are included for completeness of the thesis.

The error function $\operatorname{erf}(z)$ will be used multiple times in this appendix. It is defined as a definite integral of the normal distribution [97]:

$$\operatorname{erf}(z) \triangleq \frac{2}{\sqrt{\pi}} \int_0^z \exp(-u^2)du \tag{A.1}$$

Four key properties of the error function are [97]:

$$\operatorname{erf}(0) = 0 \tag{A.2}$$

$$\lim_{z \to +\infty} \operatorname{erf}(z) = 1 \tag{A.3}$$

$$\lim_{z \to -\infty} \operatorname{erf}(z) = -1 \tag{A.4}$$

$$\operatorname{erf}(-z) = -\operatorname{erf}(z) \tag{A.5}$$

Figure A.1 shows the error function, $\operatorname{erf}(z)$, plotted for $-3 < z < 3$. The definite integral over the normal distribution from which it is defined is also shown.

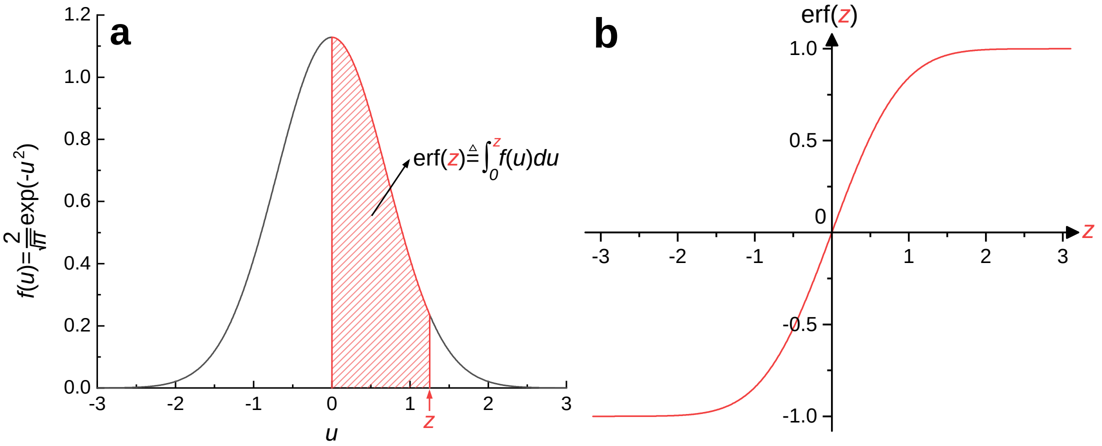

Fig. A.1: Illustration of the error function. **a**, Definition of the error function as the integral between 0 and z of the normal distribution. **b**, The error function, $\operatorname{erf}(z)$.

A.1 Gaussian Filter Impulse Response

The Gaussian filter impulse response in Chapter 4 Equation 4.13 can be derived by performing the inverse Fourier transform on the normalised Gaussian filter frequency response in Chapter 4 Equation 4.12 [95].

Starting with the normalised Gaussian filter frequency response, $\hat{H}(f)$, in terms of the cut-off frequency, f_c:

$$\hat{H}(f) = \exp\bigg(- \left(\frac{f}{f_c}\right)^2 \frac{\ln 2}{2}\bigg) \tag{A.6}$$

The impulse response is derived by performing the inverse Fourier transform of the Gaussian filter frequency response:

$$\hat{h}(t) = \mathcal{F}^{-1}\{\hat{H}(f)\} \tag{A.7}$$

The full Fourier transform is:

$$\hat{h}(t) = \int_{-\infty}^{\infty} \exp\bigg(- \left(\frac{f}{f_c}\right)^2 \frac{\ln 2}{2}\bigg) \exp(2\pi i f t) df \tag{A.8}$$

To evaluate this integral, the exponential for the impulse term is first combined with the exponential for the Gaussian filter frequency response.

$$\hat{h}(t) = \int_{-\infty}^{\infty} \exp\bigg(- \left(\frac{f}{f_c}\right)^2 \frac{\ln 2}{2} + 2\pi i f t\bigg) df \tag{A.9}$$

This is then a standard integral [97] of the form:

$$\int_{-\infty}^{\infty} \exp(-a^2x^2 \pm bx)dx = \frac{\sqrt{\pi}}{a} \exp\left(\frac{b^2}{4a^2}\right), \quad \text{for } a > 0 \tag{A.10}$$

Where the terms x, a^2 and b are defined in this context as:

$$x \triangleq f \qquad a^2 \triangleq \frac{\ln 2}{2f_c^2} \qquad b \triangleq 2\pi i t \tag{A.11}$$

Using Equation A.10 to solve Equation A.9, the result of the integral is then:

$$\hat{h}(t) = f_c \sqrt{\frac{2\pi}{\ln 2}} \exp\bigg(- \left(\frac{2\pi^2 f_c^2}{\ln 2}\right) t^2\bigg) \tag{A.12}$$

Which is the normalised Gaussian filter impulse response shown in Chapter 4 Equation 4.13.

A.2 Confirming Normalisation of the Gaussian Impulse Filter

The derived Gaussian filter impulse response in Equation A.12 is correctly normalised if its integral over all time is equal to 1 [95].

$$\int_{-\infty}^{\infty} \hat{h}(t)dt = \int_{-\infty}^{\infty} f_c\sqrt{\frac{2\pi}{\ln 2}} \exp\left(-\left(\frac{2\pi^2 f_c^2}{\ln 2}\right)t^2\right)dt \tag{A.13}$$

This integral can be evaluated by substitution using the error function defined in Equation A.1. First, we define u^2 in this context:

$$u^2 \triangleq \frac{2\pi^2 f_c^2 t^2}{\ln 2} \tag{A.14}$$

u and du are then:

$$u = \sqrt{\frac{2}{\ln 2}}\pi f_c t \qquad\qquad du = \sqrt{\frac{2}{\ln 2}}\pi f_c dt \tag{A.15}$$

If we then substitute for t and dt in Equation A.13 (the integration limits after substitution remain from $-\infty$ to ∞):

$$\begin{aligned}
\int_{-\infty}^{\infty} \hat{h}(t)dt &= \frac{\cancel{f_c}}{\pi\cancel{f_c}}\sqrt{\frac{\cancel{\ln 2}}{\cancel{2}}}\sqrt{\frac{\cancel{2}\pi}{\cancel{\ln 2}}}\int_{-\infty}^{\infty} \exp(-u^2)du \\
\int_{-\infty}^{\infty} \hat{h}(t)dt &= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} \exp(-u^2)du
\end{aligned} \tag{A.16}$$

We can split this into two separate integrals:

$$\int_{-\infty}^{\infty} \hat{h}(t)dt = \frac{1}{\sqrt{\pi}}\left(\int_{0}^{\infty} \exp(-u^2)du - \int_{0}^{-\infty} \exp(-u^2)du\right) \tag{A.17}$$

We can then evaluate these integrals using the definition of the error function in Equation A.1 and its key properties given in Equations A.2, A.3 and A.4:

$$\begin{aligned}
\int_{-\infty}^{\infty} \hat{h}(t)dt &= \frac{1}{\sqrt{\pi}}\left(\lim_{z\to+\infty}\int_{0}^{z} \exp(-u^2)du - \lim_{z\to-\infty}\int_{0}^{z} \exp(-u^2)du\right) \\
\int_{-\infty}^{\infty} \hat{h}(t)dt &= \frac{1}{\sqrt{\pi}}\left(\lim_{z\to+\infty}\left[\frac{\sqrt{\pi}}{2}\operatorname{erf}(z)\right]_0^z - \lim_{z\to-\infty}\left[\frac{\sqrt{\pi}}{2}\operatorname{erf}(z)\right]_0^z\right) \\
\int_{-\infty}^{\infty} \hat{h}(t)dt &= \frac{1}{\sqrt{\pi}}\left(\frac{\sqrt{\pi}}{2}(1-0) - \frac{\sqrt{\pi}}{2}(-1-0)\right) = \frac{\cancel{\sqrt{\pi}}}{\cancel{\sqrt{\pi}}}\left(\frac{1}{2}+\frac{1}{2}\right) \\
\int_{-\infty}^{\infty} \hat{h}(t)dt &= 1
\end{aligned} \tag{A.18}$$

The integral over all time of the filter impulse response given in Equation A.12 is equal to 1. The filter impulse response is therefore correctly normalised.

A.3 Gaussian NRZ Pulse Shape

The Gaussian NRZ pulse shape is derived by convolving the impulse response of a Gaussian filter, $\hat{f}(\phi)$, given in Chapter 4 Equation 4.21 with an ideal rectangular NRZ pulse shape, $w(\phi)$, given in Chapter 4 Equation 4.11 [95]. For convenience, this derivation performs the convolution in terms of phase, ϕ and standard deviation of the Gaussian impulse in terms of phase, k_{pls}, but it can equivalently be performed in terms of time, t and standard deviation in terms of time, σ_t, by substitution of the definitions $\phi \triangleq Bt$ and $k_{\text{pls}} \triangleq B\sigma_t$ given in Chapter 4 Equations 4.6 and 4.17, where B is the symbol rate.

The convolution to find the NRZ pulse shape is:

$$\begin{aligned} v(\phi) &\triangleq \left(\hat{h} * w\right)(\phi) \\ v(\phi) &= \int_{-\infty}^{\infty} \hat{h}(\alpha) w(\phi - \alpha) d\alpha \end{aligned} \tag{A.19}$$

Where the ideal square NRZ pulse shape, $w(\phi)$, is defined as [95]:

$$w(\phi) \triangleq \begin{cases} 1, & \text{if } |\phi| < \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \tag{A.20}$$

The convolution after substituting in the Gaussian filter impulse response in Equation A.20 and moving the constant part of the integral outside is:

$$v(\phi) = \frac{1}{k_{\text{pls}}\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{\alpha^2}{2k^2}\right) w(\phi - \alpha) d\alpha \tag{A.21}$$

We can split this integral into three parts, considering that $w(\phi - \alpha) = 1$ for $\frac{1}{2} < \phi - \alpha < \frac{1}{2}$ and $w(\phi - \alpha) = 0$ for all other values of $\phi - \alpha$:

$$\begin{aligned} v(\phi) = \frac{1}{k_{\text{pls}}\sqrt{2\pi}} \Bigg(&\int_{\phi+\frac{1}{2}}^{\infty} \exp\left(-\frac{\alpha^2}{2k_{\text{pls}}^2}\right) \cancelto{0}{w(\phi - \alpha)} d\alpha \\ &+ \int_{\phi-\frac{1}{2}}^{\phi+\frac{1}{2}} \exp\left(-\frac{\alpha^2}{2k_{\text{pls}}^2}\right) \cancelto{1}{w(\phi - \alpha)} d\alpha \\ &+ \int_{-\infty}^{\phi-\frac{1}{2}} \exp\left(-\frac{\alpha^2}{2k_{\text{pls}}^2}\right) \cancelto{0}{w(\phi - \alpha)} d\alpha \Bigg) \end{aligned} \tag{A.22}$$

In the first and third integrals, $w(\phi-\alpha)$ is equal to 0 over the entirety of the integral. In second integral that remains, $w(\phi-\alpha)$ is equal to 1. The remaining part of the convolution is then:

$$v(\phi)=\frac{1}{k_{\text{pls}}\sqrt{2\pi}}\int_{\phi-\frac{1}{2}}^{\phi+\frac{1}{2}}\exp\left(-\frac{\alpha^2}{2k_{\text{pls}}^2}\right)d\alpha \tag{A.23}$$

This integral can be evaluated by substitution using the error function defined in Equation A.1. If we first define u^2 in this context:

$$u^2\triangleq\frac{\alpha^2}{2k_{\text{pls}}^2} \tag{A.24}$$

u and du are then:

$$u=\frac{\alpha}{\sqrt{2}k_{\text{pls}}}\qquad\qquad du=\frac{d\alpha}{\sqrt{2}k_{\text{pls}}} \tag{A.25}$$

If we then substitute for α and $d\alpha$ in Equation A.13, then the integral becomes:

$$v(\phi)=\frac{1}{\sqrt{\pi}}\int_{\frac{1}{\sqrt{2}k_{\text{pls}}}(\phi-\frac{1}{2})}^{\frac{1}{\sqrt{2}k_{\text{pls}}}(\phi+\frac{1}{2})}\exp(-u^2)du \tag{A.26}$$

This can then be split into two integrals:

$$v(\phi)=\frac{1}{\sqrt{\pi}}\int_{0}^{\frac{1}{\sqrt{2}k_{\text{pls}}}(\phi+\frac{1}{2})}\exp(-u^2)du-\frac{1}{\sqrt{\pi}}\int_{0}^{\frac{1}{\sqrt{2}k_{\text{pls}}}(\phi-\frac{1}{2})}\exp(-u^2)du \tag{A.27}$$

Each of these integrals can then be evaluated using the definition of the error function in Equation A.1 and

$$\begin{aligned}v(\phi)&=\left[\frac{1}{2}\,\text{erf}(u)\right]_0^{\frac{1}{\sqrt{2}k_{\text{pls}}}(\phi+\frac{1}{2})}-\left[\frac{1}{2}\,\text{erf}(u)\right]_0^{\frac{1}{\sqrt{2}k_{\text{pls}}}(\phi-\frac{1}{2})}\\ v(\phi)&=\frac{1}{2}\Bigg(\bigg(\text{erf}\left(\frac{\phi+\frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)-\text{erf}(0)\bigg)-\bigg(\text{erf}\left(\frac{\phi-\frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)-\text{erf}(0)\bigg)\Bigg)\end{aligned} \tag{A.28}$$

Finally, we use the property in Equation A.2 that $\text{erf}(0)=0$:

$$v(\phi)=\frac{1}{2}\Bigg(\text{erf}\left(\frac{\phi+\frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)-\text{erf}\left(\frac{\phi-\frac{1}{2}}{\sqrt{2}k_{\text{pls}}}\right)\Bigg) \tag{A.29}$$

Equation A.29 is the end result of the convolution and is the Gaussian NRZ pulse shape given in Chapter 4 Equation 4.25.

A.4 Clock Jittered Alexander Phase Detector Phase Error Expectation

An Alexander phase detector can be defined as having the following ideal phase error, $\Phi_{\text{ideal}}(\phi)$, depending on the phase offset in symbols, ϕ, between the sampling clock and the incoming embedded clock within the sampled data. Unlike typical definitions of the clock phase error for an ideal phase error, such as in [102], that are defined only for $-\frac{1}{2} < \phi < \frac{1}{2}$, this definition includes the repetition of the ideal phase response at values of ϕ beyond $-\frac{1}{2} < \phi < \frac{1}{2}$:

$$\Phi_{\text{ideal}}(\phi) \triangleq \begin{cases} -1, & \text{if } n - \frac{1}{2} < \phi < n \\ 0, & \text{if } \phi = \frac{1}{2}n \\ +1, & \text{if } n < \phi < n + \frac{1}{2} \end{cases} \tag{A.30}$$

where $n \in \mathbb{Z}$

The expectation of the sampling position, $\text{E}[\Phi_{\text{jittered}}(\phi)]$, as a function of mean clock phase offset, ϕ, after accounting for jitter between the sampling clock and the embedded clock within the data can be modelled by performing a convolution between the PDF of the jitter, PDF_{jit}, and the ideal phase error, $\Phi_{\text{ideal}}(\phi)$ [126].

$$\text{E}[\Phi_{\text{jittered}}(\phi)] = (\text{PDF}_{\text{jit}} * \Phi_{\text{ideal}})\,(\phi) \tag{A.31}$$

The PDF of the jitter between the sampling clock and the incoming embedded clock within the sampled data can be modelled by a Gaussian, PDF_{jit} with a mean of 0 symbols and a standard deviation of k_{jit}:

$$\text{PDF}_{\text{jit}}(\phi) = \frac{1}{k_{\text{jit}}\sqrt{2\pi}} \exp\left(-\frac{\phi^2}{2k_{\text{jit}}^2}\right) \tag{A.32}$$

The full convolution is then:

$$\text{E}[\Phi_{\text{jittered}}(\phi)] = \int_{-\infty}^{\infty} \frac{1}{k_{\text{jit}}\sqrt{2\pi}} \exp\left(-\frac{\alpha^2}{2k_{\text{jit}}^2}\right) \Phi_{\text{ideal}}(\phi - \alpha)\, d\alpha \tag{A.33}$$

This integral can be split into an infinite sum of two series of integrals, of positive integrals where $\Phi_{\text{ideal}}(\phi - \alpha) = 1$, for $2n < \phi - \alpha < n + \frac{1}{2}$ (where $n \in \mathbb{Z}$), and of negative integrals where $\Phi_{\text{ideal}}(\phi - \alpha) = -1$, for $n - \frac{1}{2} < \phi - \alpha < n$ (where $n \in \mathbb{Z}$).

$$\begin{aligned} \mathrm{E}[\Phi_{\text{jittered}}(\phi)] = \frac{1}{k_{\text{jit}}\sqrt{2\pi}} \Bigg(\sum_{n=-\infty}^{\infty} \Bigg(\int_{\phi-n-\frac{1}{2}}^{\phi-n} \exp\Bigg(-\frac{\alpha^2}{2k_{\text{jit}}^2}\Bigg) d\alpha \\ - \int_{\phi-n}^{\phi-n+\frac{1}{2}} \exp\Bigg(-\frac{\alpha^2}{2k_{\text{jit}}^2}\Bigg) d\alpha \Bigg)\Bigg) \end{aligned} \tag{A.34}$$

These integrals can be solved by substitution using the definition of the error function in Equation A.1. If we first define u^2 in this context:

$$u^2 \triangleq \frac{\alpha^2}{2k_{\text{jit}}^2} \tag{A.35}$$

u and du are then:

$$u = \frac{\alpha}{\sqrt{2}k_{\text{jit}}} \qquad du = \frac{d\alpha}{\sqrt{2}k_{\text{jit}}} \tag{A.36}$$

The integrals after substitution then become:

$$\begin{aligned} \mathrm{E}[\Phi_{\text{jittered}}(\phi)] = \frac{1}{\sqrt{\pi}} \Bigg(\sum_{n=-\infty}^{\infty} \Bigg(\int_{\frac{1}{\sqrt{2}k_{\text{jit}}}(\phi-n-\frac{1}{2})}^{\frac{1}{\sqrt{2}k_{\text{jit}}}(\phi-n)} \exp(-u^2) du \\ - \int_{\frac{1}{\sqrt{2}k_{\text{jit}}}(\phi-n)}^{\frac{1}{\sqrt{2}k_{\text{jit}}}(\phi-n+\frac{1}{2})} \exp(-u^2) du \Bigg)\Bigg) \end{aligned} \tag{A.37}$$

We can then split these two integrals into four integrals:

$$\begin{aligned} \mathrm{E}[\Phi_{\text{jittered}}(\phi)] = \frac{1}{\sqrt{\pi}} \Bigg(\sum_{n=-\infty}^{\infty} \Bigg(\int_{0}^{\frac{1}{\sqrt{2}k_{\text{jit}}}(\phi-n)} \exp(-u^2) du \\ - \int_{0}^{\frac{1}{\sqrt{2}k_{\text{jit}}}(\phi-n-\frac{1}{2})} \exp(-u^2) du \\ - \int_{0}^{\frac{1}{\sqrt{2}k_{\text{jit}}}(\phi-n+\frac{1}{2})} \exp(-u^2) du \\ + \int_{0}^{\frac{1}{\sqrt{2}k_{\text{jit}}}(\phi-n)} \exp(-u^2) du \Bigg)\Bigg) \end{aligned} \tag{A.38}$$

We can then evaluate these integrals using the definition of the error function:

$$\begin{aligned}\mathrm{E}[\Phi_{\mathrm{jittered}}(\phi)] = \frac{1}{2}\Bigg(\sum_{n=-\infty}^{\infty} \Bigg(\operatorname{erf}\left(\frac{\phi - n}{\sqrt{2}k_{\mathrm{jit}}}\right) - \operatorname{erf}\left(\frac{\phi - n - \frac{1}{2})}{\sqrt{2}k_{\mathrm{jit}}}\right) \\ - \operatorname{erf}\left(\frac{\phi - n + \frac{1}{2}}{\sqrt{2}k_{\mathrm{jit}}}\right) + \operatorname{erf}\left(\frac{\phi - n}{\sqrt{2}k_{\mathrm{jit}}}\right)\Bigg)\Bigg) \qquad \text{(A.39)}\end{aligned}$$

Finally, if we observe that the $\phi - n - \frac{1}{2}$ term is just shifted by $n = 1$ from $\phi - n + \frac{1}{2}$, we can gather positive and negative terms:

$$\mathrm{E}[\Phi_{\mathrm{jittered}}(\phi)] = \sum_{n=-\infty}^{\infty} \left(\operatorname{erf}\left(\frac{\phi - n}{\sqrt{2}k_{\mathrm{jit}}}\right) - \operatorname{erf}\left(\frac{\phi - n - \frac{1}{2}}{\sqrt{2}k_{\mathrm{jit}}}\right)\right) \qquad \text{(A.40)}$$

Equation A.40 gives the end exact result of the convolution, for any value of k_{jit}. However, in a practical communication system the jitter standard deviation, k_{jit}, needs to be much smaller than one symbol, i.e. $k_{\mathrm{jit}} \ll 1$, to minimise the large increase in error probability, p_e, associated with large k_{jit}. For $k_{\mathrm{jit}} < \frac{1}{8\sqrt{2}}$, looking at, for example, the centre-most positive term where $n = 0$, $\operatorname{erf}\left(\frac{\phi}{\sqrt{2}k_{\mathrm{jit}}}\right)$ reaches at least 99.5% of its asymptotic value of +1 and -1 by +$\frac{1}{4}$ symbols and -$\frac{1}{4}$ symbols respectively. $k_{\mathrm{jit}} = \frac{1}{8\sqrt{2}} \approx 0.0884$ symbols is very large, and would lead to complete eye closure due to error probability increase as shown in Chapter 4 Figure 4.19.

Consequently, we can reasonably assume that $k_{\mathrm{jit}} < \frac{1}{8\sqrt{2}}$. Under this condition, we can simplify the result of the convolution, approximating that:

$$\operatorname{erf}\left(\frac{\phi - n}{\sqrt{2}k_{\mathrm{jit}}}\right) \approx \begin{cases} -1, & \text{if } \phi - n < -\frac{1}{4} \\ \operatorname{erf}\left(\frac{\phi - n}{\sqrt{2}k_{\mathrm{jit}}}\right), & \text{if } -\frac{1}{4} < \phi - n < \frac{1}{4} \\ +1, & \text{if } \phi - n > \frac{1}{4} \end{cases} \qquad \text{(A.41)}$$

where $n \in \mathbb{Z}$ and $k_{\mathrm{jit}} < \frac{1}{8\sqrt{2}}$

$$\operatorname{erf}\left(\frac{\phi - n - \frac{1}{2}}{\sqrt{2}k_{\mathrm{jit}}}\right) \approx \begin{cases} -1, & \text{if } \phi - n - \frac{1}{2} < -\frac{1}{4} \\ \operatorname{erf}\left(\frac{\phi - n - \frac{1}{2}}{\sqrt{2}k_{\mathrm{jit}}}\right), & \text{if } -\frac{1}{4} < \phi - n - \frac{1}{2} < \frac{1}{4} \\ +1, & \text{if } \phi - n - \frac{1}{2} > \frac{1}{4} \end{cases} \qquad \text{(A.42)}$$

where $n \in \mathbb{Z}$ and $k_{\mathrm{jit}} < \frac{1}{8\sqrt{2}}$

Under the assumption that $k_{\text{jit}} < \frac{1}{8\sqrt{2}}$, the contributions of each $\operatorname{erf}(z)$ term that are more than $\frac{1}{4}$ symbols away from ϕ can be ignored, as they cancel in pairs, leaving a single $\operatorname{erf}(z)$ or $-\operatorname{erf}(z)$ contribution to the expectation of the clock phase error.

As an example, if we look specifically at the region where the phase of the sampling clock is within $\frac{1}{4}$ symbol of the clock embedded within the sampled data ($-\frac{1}{4} < \phi < \frac{1}{4}$, $n = 0$), all contributions other than $\operatorname{erf}\left(\frac{\phi}{\sqrt{2}k_{\text{jit}}}\right)$ cancel in pairs:

$$\begin{aligned}
\mathrm{E}[\Phi_{\text{jittered}}(\phi)] &= \operatorname{erf}\left(\frac{\phi}{\sqrt{2}k_{\text{jit}}}\right) + \cancelto{-1}{\operatorname{erf}\left(\frac{\phi-1}{\sqrt{2}k_{\text{jit}}}\right)} + \cancelto{1}{\operatorname{erf}\left(\frac{\phi+1}{\sqrt{2}k_{\text{jit}}}\right)} \\
&\quad - \cancelto{-1}{\operatorname{erf}\left(\frac{\phi-\frac{1}{2}}{\sqrt{2}k_{\text{jit}}}\right)} - \cancelto{1}{\operatorname{erf}\left(\frac{\phi+\frac{1}{2}}{\sqrt{2}k_{\text{jit}}}\right)} \\
&\quad + \ldots + \cancelto{-1}{\operatorname{erf}\left(\frac{\phi-n}{\sqrt{2}k_{\text{jit}}}\right)} + \cancelto{1}{\operatorname{erf}\left(\frac{\phi+n}{\sqrt{2}k_{\text{jit}}}\right)} \\
&\quad - \cancelto{-1}{\operatorname{erf}\left(\frac{\phi-n-\frac{1}{2}}{\sqrt{2}k_{\text{jit}}}\right)} - \cancelto{1}{\operatorname{erf}\left(\frac{\phi+n-\frac{1}{2}}{\sqrt{2}k_{\text{jit}}}\right)} \\
&\approx \operatorname{erf}\left(\frac{\phi}{\sqrt{2}k_{\text{jit}}}\right) \qquad \text{(A.43)}
\end{aligned}$$

where $n \in \mathbb{Z}$, $k_{\text{jit}} < \frac{1}{8\sqrt{2}}$ and $-\frac{1}{4} < \phi < \frac{1}{4}$.

We can generalise this to any clock phase offset, ϕ, by performing the same pair cancellation process at all values of ϕ:

$$\mathrm{E}[\Phi_{\text{jittered}}(\phi)] \approx \begin{cases} \operatorname{erf}\left(\frac{\phi-n}{\sqrt{2}k_{\text{jit}}}\right), & \text{if } n-\frac{1}{4} < \phi < n+\frac{1}{4} \\ -\operatorname{erf}\left(\frac{\phi-n-\frac{1}{2}}{\sqrt{2}k_{\text{jit}}}\right), & \text{if } n+\frac{1}{4} < \phi < n+\frac{3}{4} \end{cases} \qquad \text{(A.44)}$$

where $n \in \mathbb{Z}$ and $k_{\text{jit}} < \frac{1}{8\sqrt{2}}$

Equation A.44 (also shown in Chapter 5 Equation 5.3) therefore gives a very good approximation to $\mathrm{E}[\Phi_{\text{jittered}}(\phi)]$ for $k_{\text{jit}} < \frac{1}{8\sqrt{2}}$.

A.5 Contribution Statement

K.A.C. performed all mathematical derivations, supervised by Z.L.. E.V.V. provided helpful advice through multiple discussions.

Appendix B

Clock Synchronisation Assisted Clock and Data Recovery FPGA Hardware Design

This appendix describes the FPGA hardware designed and implemented to demonstrate CSA-CDR in a real-time optically-switched system in Chapters 6 and 7, a prototype of the proposed CSA-CDR architecture illustrated and described in Chapter 3 Section 3.4.

B.1 Overview of FPGA Hardware Design

Figure B.1 shows an overview of the FPGA design used to implement CSA-CDR. If required, this design performs clock phase measurements and sends clock phase updates according to the principle of operation shown in Chapter 6 Figure 6.1 and also counts errors in received data packets. This same FPGA design is replicated across the three FPGA nodes in Chapter 6 Figure 6.8 and Chapter 7 Figure 7.5, and can be configured to operate as a transmitting node or a receiving node. In the figure: orange boxes represent hardware modules used to implement CSA-CDR, blue boxes represent hardware components used to transmit and receive data packets, and black boxes represent hardware modules used to output error counts and clock phase updates to an external MATLAB controller; solid black arrows with regular font labels show signals used to communicate between hardware modules, solid green arrows with regular font labels show clock signals, dashed black arrows with italic font labels show signals used to configure hardware components; and dashed grey lines surrounding areas of the design, with grey italic labels, show the three different main clock domains in the design (the 400 MHz transmitter and receiver parallel clocks, and the 100 MHz system clock). Some signals between hardware modules in Figure B.1 also include enable and FPGA hardware address signals as well as the information being transmitted (e.g. for communication with first-in first-out queues (FIFOs)). All modules within a clock domain are driven by that domain's associated clock signal. Clock inputs for each module are not shown to minimise illustrative complexity, and note that either side of the FIFOs that sit between two clock domains are driven by each clock domain's clock signal (since they are used to transmit data between different clock domains).

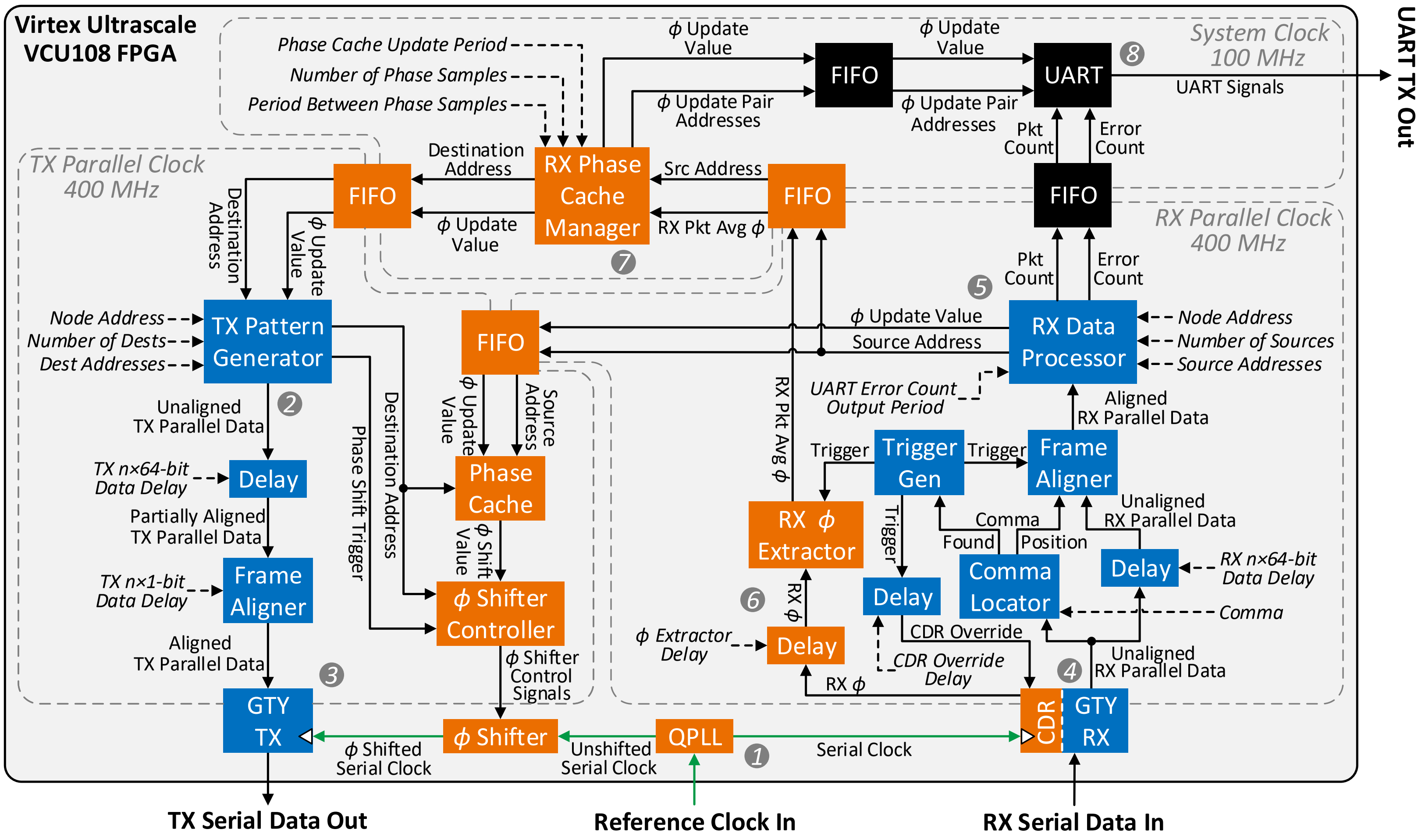

Fig. B.1: CSA-CDR FPGA hardware used in Chapter 6 and 7. CDR, clock and data recovery; QPLL, quad phase lock loop; TX, transmitter; RX, receiver; src, source; dest(s), destination(s); ϕ, phase; avg, average; pkt, packet; PRBS, pseudo-random binary sequence; FIFO, first in first out; UART, Universal Asynchronous Receiver/Transmitter. Reference clock frequency: 800 MHz; serial clock frequency: 12.8 GHz (half-rate).

The FPGA hardware shown in Figure B.1 used to perform CSA-CDR functions as follows (numbered labels correspond to their twins in Figure B.1):

➊ **Serial clock generation and distribution.** The 800 MHz reference clock used to frequency synchronise the FPGA drives a QPLL, which is based on an LC-tank oscillator. The QPLL generates a 12.8 GHz (half-rate) clock that is used to drive the serial clock of the transmitter and receiver sides of a Xilinx UltraScale GTY FPGA transceiver, can and also be used to provide a reference clock for three other adjacent Xilinx UltraScale GTY FPGA transceivers (hence quad PLL). The phase of the transmitter half-rate clock is shifted by a digital PI-based clock phase shifter integrated in the FPGA. The phase of the receiver half-rate clock is shifted by a digital PI-based CDR. The transmitter and receiver generate the transmitter and receiver 400 MHz parallel clocks respectively by dividing the clock phase shifted half-rate clocks by 32. The 100 MHz system clock, which is not synchronous with the transmitter and receiver parallel clocks, is generated from an external 400 MHz oscillator on each FPGA, which is used to drive hardware that configures that FPGA transceivers and the configuration parameters shown for the hardware in Figure B.1.

➋ **Transmitter packet generation, delay and frame alignment.** Data packets are generated by a TX pattern generator, in 64 bit parallel format. These packets contain, in order, a data payload formed of 2×512 bit De Bruijn sequences [107] of 2^9 bits length (511 bit PRBS-9 sequence with a polynomial of x^9+x^5+1 [108] with a final appended 0 bit), a 16 bit comma sequence for frame alignment, a 30 bit header and a final 466 bit De Bruijn sequence to pad the remainder of the packet. The header contains, in order, a 4 bit source address, a 4 bit destination address, an 8 bit clock phase update value (1 bit sign followed by 7 bits magnitude, all 8 bits 0 if not required), and an 8 bit cyclic redundancy check (CRC) to check for data corruption. The entire header is encoded to a length of 30 bits using an 8B/10B encoder [127], with the running disparity set to always be 0. After generation, the data packets then pass through a $n\times64$ bit delay ring-buffer, followed by a frame aligner that further delays the packets by 0 to 63 bits with a barrel shifter. Both the transmitter-side delay and frame aligner are manually configured once at system startup to align packets leaving the two transmitter FPGAs, a task which would be performed automatically in a full commercial implementation.

➌ **Transmitter serialisation and packet clock phase shift.** The 64 bit parallel format data packets are then serialised to 25.6 Gb/s serial format by the transmitter side of a Xilinx UltraScale GTY FPGA transceiver using a PISO (integrated within the FPGA transceiver), before being transmitted from the FPGA. The clock phase of packets leaving the transmitter are shifted on a per-packet basis by changing the clock phase of the transmitter serial clock with a PI-based clock phase shifter, which is controlled by a clock phase shifter controller. The clock phase shifter controller in-turn receives the correct clock phase values from the clock phase cache, which outputs a clock phase shift output value based on which destination address control signal it receives from the TX pattern generator (which matches the destination address embedded in generated packets).

❹ **Packet reception; receiver deserialisation, delay and frame alignment.** Serial data packets first arrive at 25.6 Gb/s at the receiver side of a Xilinx UltraScale GTY FPGA transceiver. A PI-based Xilinx CDR circuit locks to the clock phase of incoming packets, and outputs a receiver clock phase value at a frequency of 400 MHz as well as a phase shifted version of the half-rate clock. The half-rate clock is used to sample the incoming packets, which are then deserialised to 64 bit parallel format data packets with a serial-in parallel-out (SIPO) (integrated within the FPGA transceiver). A comma locator is used to find the position of the comma in the incoming parallel data (from bit 0 to 63), and to generate a signal indicating that the comma has been found. The comma locator looks for an exact match between the comma and the expected sequence. The found signal from the comma locator drives a trigger generator, which in-turn generates three triggers that drive the frame aligner, CDR override delay ring buffer (the output of which is configured to reset the RX CDR phase between incoming packets) and the receiver clock phase extractor. The incoming data packets are delayed by a $n\times64$ bit delay ring-buffer, by the length of the data payload prior to the comma, such that the first bits of the data payload of each packet arrive at the frame aligner immediately after the comma sequence is found. A receiver frame aligner then aligns the first bit of the comma with the first bit of the 64 bit parallel bus using a barrel shifter.

❺ **Receiver packet BER measurement and header decoding.** The aligned 64 bit parallel format data packets are processed by an RX data processor. This hardware module contains a PRBS sequence error detector, which compares the incoming data payload with a reference sequence to count bit errors that fall in 64×16 bit bin intervals in the data payload; a CRC error detector, which checks for data corruption in the packet header; an 8B/10B decoder [127] to decode the packet header (with the running disparity set to always be 0); and a label extractor, which extracts the header information (source address, destination address and transmitter clock phase update (if required)) from the packet header. With a configurable repetition period, the error counts from the 64×16 bit bins are sent to the universal asynchronous receiver transmitter (UART) controller through a clock domain crossing FIFO, along with the number of packets counted. If a transmitter clock phase update is required, this is sent to the transmitter clock phase cache (to replace the previous value) through a clock domain crossing FIFO along with the packet source address for which the clock phase update is required.

❻ **Receiver clock phase measurement.** The clock phase of the second and third packet sequences within each incoming packet is averaged by a receiver clock phase extractor (performed using bit shift operations to minimise hardware complexity). Clock phase values from the receiver CDR arriving at this hardware module are delayed such that the averaged clock phase value output from this module is synchronised with the extraction of the packet source address from the packet header within the RX data processor. The average packet clock phase values along with their associated source addresses are passed to the RX clock phase cache manager through a clock domain crossing FIFO. The RX clock phase cache manager averages clock phase values across 2^n packets from each source address (again performed using bit shift operations to minimise hardware complexity), with a configurable period between sampled packets and a configurable value for n.

❼ **Clock phase updating.** Once 2^n packets from a source address have been sampled and averaged, a clock phase update is passed to the TX pattern generator through a clock domain crossing FIFO. This clock phase update includes the clock phase offset value measured and the destination address for the remote node clock phase cache to be updated (which is the same address as the source address for which clock phase of the 2^n packets were measured). The TX pattern generator waits until a packet with the correct destination address being generated, at which point it embeds the clock phase update in the outgoing packet. The RX clock phase cache manager also sends the clock phase update direction and magnitude, as well as the phase update pair addresses (the FPGA board address / source address and the destination address), to the UART controller through a FIFO. This FIFO, unlike the other FIFOs in the design, is used to buffer phase updates until the UART controller is free to transmit them rather than to cross between different clock domains.

❽ **UART output.** The UART controller periodically outputs bit error counts detected in received data packets along with the number of data packets counted, and also outputs information regarding each clock phase update, which includes the addresses of the pair of nodes for which the clock phase update has occurred as well as the direction and magnitude of the clock phase update. The information that is output over UART is received by an instance of Xilinx Vivado, which is executed within a MATLAB script. This script records receiver-measured PRBS errors falling in each of the 16×64 bins and tracks clock phase updates being requested by the receiver FPGA. The script is also used to automate the process of experimental result acquisition and experiment control.

Miscellaneous hardware modules. Additional hardware modules are also present in the FPGA design but are not shown in Figure B.1 to minimise the complexity of the illustration (because they perform operations associated with debug and system control, and are not used for the key functions of packet transmission, packet reception, error counting, phase updating or communication over UART). These modules are: circuitry used to configure and reset the Xilinx UltraScale GTY transceiver; a counter to look for missed comma sequences (which could occur due to errors within the comma sequence); a module to count header CRC errors; virtual input/outputs (VIOs) to configure the parameters shown in Figure B.1 (shown in italics and with dashed arrows), to control the Xilinx UltraScale GTY transceiver and to provide a secondary method of monitoring error counter, via a Joint Test Action Group (JTAG) interface with Xilinx Vivado; an integrated logic analyser (ILA) to monitor the receiver CDR clock phase output values and the aligned parallel data, via a JTAG interface with Xilinx Vivado; a module to align automatically correctly align the receiver CDR override with the interpacket gap; a module to acquire statistical eyescans from the receiver side of the Xilinx UltraScale GTY transceiver and a module to store up to approximately 10 seconds of receiver clock phase values in Double Data Rate 3 Synchronous Dynamic Random-Access Memory (DDR3 SDRAM) external memory for measurement of low-frequency jitter.

B.2 Contribution Statement

K.A.C. designed, implemented and tested all FPGA hardware code. G.Z. provided helpful discussion that confirmed that the approach chosen by K.A.C. to implement CSA-CDR was appropriate.

Appendix C

Photographs

This appendix presents two photographs of the experimental setup used to acquire the results presented in this thesis. Figure C.1 shows the front of the experimental setup and Figure C.2 shows the rear of the experimental setup. The optical fibre (SMF-28 in Chapter 6 and HC-PBGF in Chapter 7) used in these experiments was contained within a thermally controlled chamber, which is out of frame of these photographs.

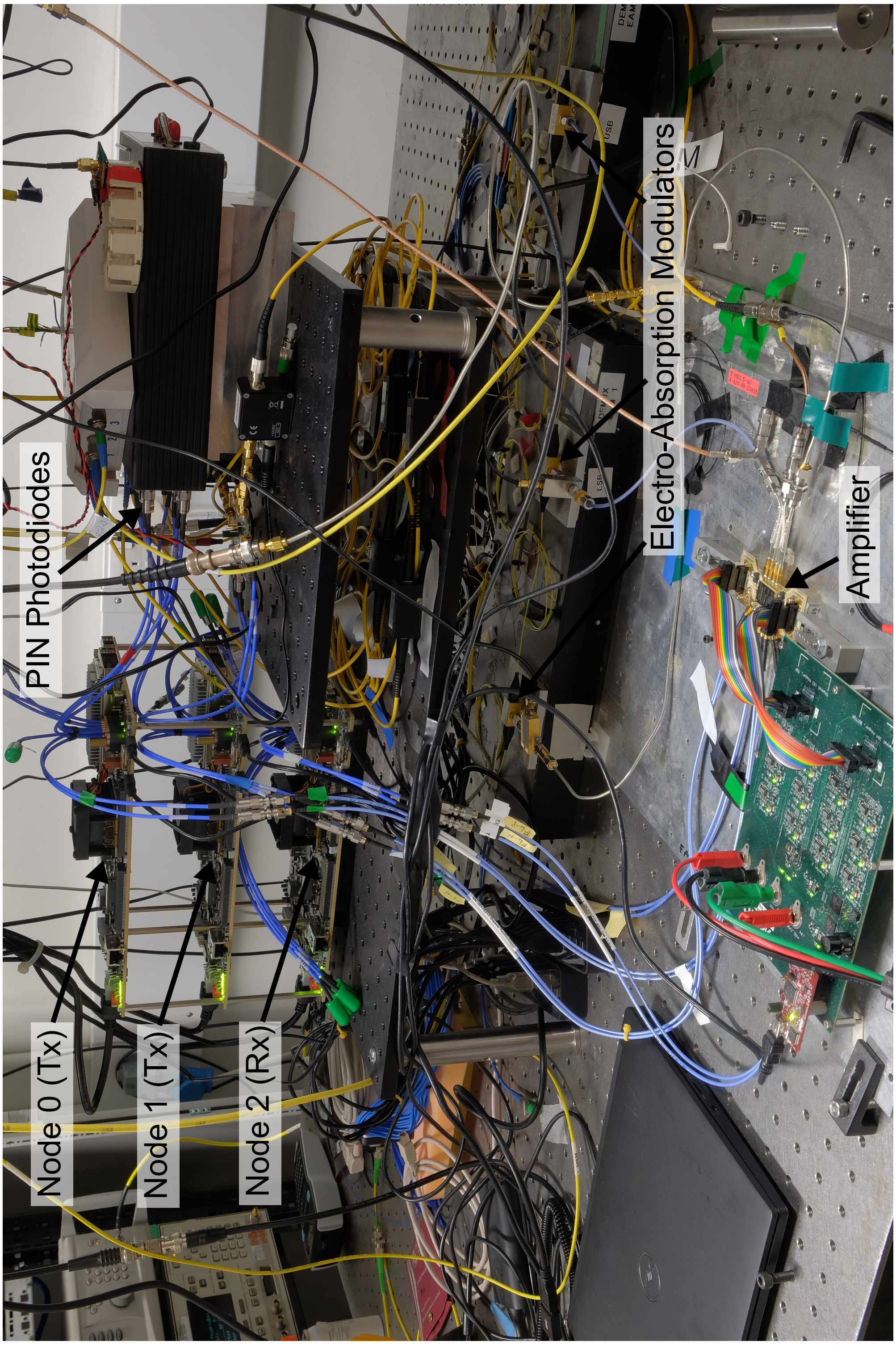

Fig. C.1: Photograph of the front of the experimental setup used to demonstrate CSA-CDR in Chapters 6 and 7. The optical fibre used in the experimental setup (SMF-28 in Chapter 6 and HC-PBGF in Chapter 7) is contained in a thermally controlled chamber, which is out of frame.

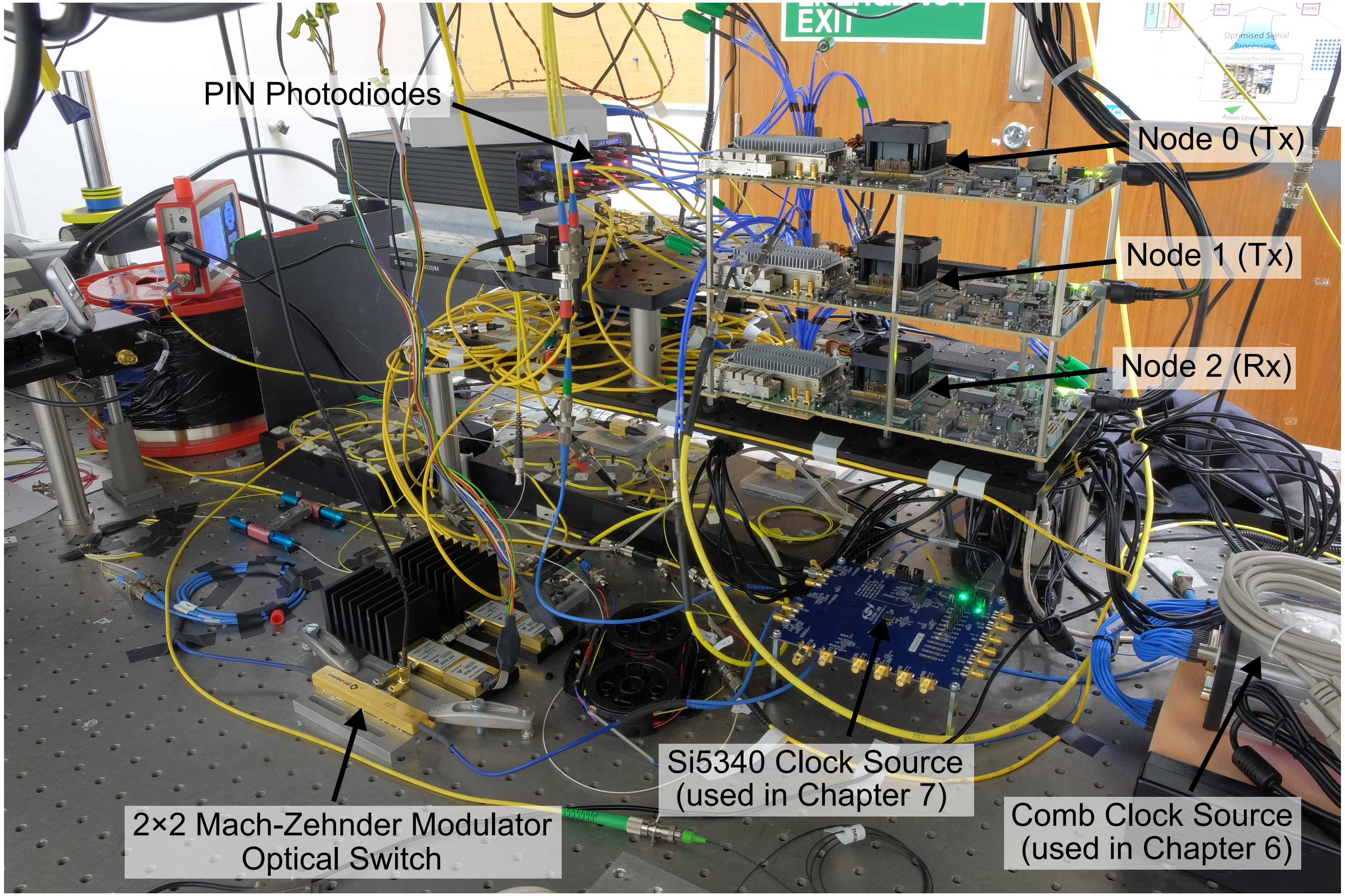
PIN Photodiodes
Node 0 (Tx)
Node 1 (Tx)
Node 2 (Rx)
2×2 Mach-Zehnder Modulator Optical Switch
Si5340 Clock Source (used in Chapter 7)
Comb Clock Source (used in Chapter 6)

Fig. C.2: Photograph of the rear of the experimental setup used to demonstrate CSA-CDR in Chapters 6 and 7. The optical fibre used in this experimental setup (SMF-28 in Chapter 6 and HC-PBGF in Chapter 7) is contained in a thermally controlled chamber, which is out of frame.