

RETHINK big: European Roadmap for Hardware and Networking Optimizations for Big Data

Gina Alioto, Paul Carpenter,
Adrián Cristal, Osman Unsal
Barcelona Supercomputing Center
Barcelona, Spain

Marcus Leich
Technische Universität Berlin
Berlin, Germany

Christophe Avare
Thales Research & Technology
Palaiseau cedex France

Abstract—This paper discusses the results of the RETHINK big Project, a 2-year Collaborative Support Action funded by the European Commission in order to write the European Roadmap for Hardware and Networking optimizations for Big Data. This industry-driven project was led by the Barcelona Supercomputing Center (BSC), and it included large industry partners, SMEs and academia. The roadmap identifies business opportunities from 89 in-depth interviews with 70 European industry stakeholders in the area of Big Data and predicts the future technologies that will disrupt the state of the art in Big Data processing in terms of hardware and networking optimizations. Moreover, it presents coordinated technology development recommendations (focused on optimizations in networking and hardware) that would be in the best interest of European Big Data companies to undertake in concert as a matter of competitive advantage.

Keywords—roadmap, Big Data, network, hardware

I. INTRODUCTION

Big Data is a large and diverse field [1] [2] [3] which is characterized by large data volumes, velocities and varieties, with standard and traditional methods and architectures unable to process and store the data within a reasonable timeframe. Processing and storage bottlenecks are leading to the adoption of specialized Big Data-optimized hardware and networking technologies, especially by the major Big Data players. For instance, Microsoft has employed FPGA acceleration to “meet the needs of Datacenter workloads [that] demand high computational capabilities, flexibility, power efficiency, and low cost” [4], resulting in a 29% reduction in tail latency for its proprietary Bing search engine. In 2015, Microsoft took Big Data hardware optimization a step further by employing FPGAs for “various kinds of deep learning approaches” [5] and it has installed FPGAs in every Azure cloud server worldwide, in order to create a cloud for artificial intelligence [6]. In June 2016, Nvidia announced plans to “tap into the big data business [which] will be a trillion-dollar business over the next few years” [7]. The deep learning hardware platform featuring GPU-accelerated training combined with ASIC-accelerated gameplay was critical to Google DeepMind’s AlphaGo program, which beat the European Go champion in tournament conditions [8]. Nvidia is also pushing GPUs in the DRIVE PX on-board compute platform for computer vision and deep learning for Advanced Driver Assistance Systems (ADAS) [9]. Facebook is focused on the OpenCompute open hardware initiative [10] [11], and it is decoupling the software

from the hardware in the network switch business, while at the same time moving toward disaggregated network architecture. Overall, there is a trend towards novel hardware for Big Data-related optimization at the large hyperscalers, as they are often the first to see the problem as well as the first to solve it.

This paper discusses the results of the RETHINK big Project, a 2-year Collaborative Support Action funded by the European Commission in order to write the European Roadmap for Hardware and Networking optimizations for Big Data. Section II describes the RETHINK big Project and methodology and Section III describes how the RETHINK big roadmap fits into other related initiatives in the European context. Section IV discusses the roadmap contributions relating to network architecture, node architecture and low-level software support, and Section V summarizes the roadmap’s key findings and recommendations. More detail is found in the complete RETHINK big roadmap [12]. Finally, Section VI concludes the paper.

II. RETHINK BIG PROJECT

The European Union, recognizing the trends of international Big Data players and in an effort to get the best return on their investment in companies in the area of Big Data analytics, called for a Roadmap to provide a coordinated set of technology development recommendations (focused on optimizations in networking and hardware) that would be in the best interest of European Big Data companies to undertake in concert as a matter of competitive advantage.

The production of this roadmap has been funded as a project, RETHINK big EC-GA No. 619788, which was led by the Barcelona Supercomputing Center and included partners from large industry to SME to academia, as shown in Table 1. In this roadmap, we identified business opportunities from European industry stakeholders in the area of Big Data and predicted future technologies that will disrupt the state of the art in Big Data processing in terms of hardware and networking optimizations. We then identified a critical mass of these stakeholders that see a clear competitive advantage enabled by embracing specific future technologies. Finally, we developed recommendations for the European Commission that will ultimately facilitate timely European industry access to these future technologies.

Partner Name	Expertise
Barcelona Supercomputing Center (BSC)	Computer architecture and system architecture
Technische Universitat Berlin (TUB)	Database systems and information management
École Polytechnique Fédérale de Lausanne (EPFL)	Database systems and applications
Centrum Voor Wiskunde en Informatica (CWI)	Hardware-conscious database technologies
University of Manchester (UoM)	Computer architecture
Universidad Politécnica de Madrid (UPM)	Data mining and warehousing
ARM Ltd. (ARM)	Silicon IP provider
Internet Memory Research (IMR)	Web-scale sourcing platform for business intelligence
Thales SA (THALES)	Situation and decision analysis, planning and optimization

Table 1: RETHINK big Project Consortium

III. A EUROPEAN ROADMAP

The RETHINK big roadmap is one piece of the framework of roadmaps (Figure 1) being put together for the European Commission. While compiling the roadmap, it was determined that many Big Data compute problems were instances of compute problems linked to the ending of Moore’s law and Dennard scaling, and beyond. As such, the roadmap scope is limited to problems with a direct impact on Big Data European industry, and general compute problems are handled by the European Technology Platform (ETP) Roadmaps (NEM, NESSI, EPoSS and Photonics21). The roadmap was developed with consideration of High Performance Computing (HPC), since HPC also routinely uses extremely large datasets [13]. Our roadmap, however, is limited to activities with a clear benefit for EU Big Data Industry, as HPC-related aspects are covered by ETP4HPC. The same is true for so-called Internet of Things on its way to becoming the Internet of Everything. The key to this nascent compute area seems to be the data itself, and we believe that the opportunities provided by IoT will be “enabled by and dependent on the tremendous data collections and compute capacities in the back-end machines and datacenters that use such data” [14]. As such, we maintain our focus on these back-end machines and data centers, leaving all other aspects to be covered under the Alliance of Internet of Things Initiative and the regulation and standards for communication at the network level under the work of the 5G-PPP (Public Private Partnership). Finally, while no discussion regarding Big Data is complete without discussion of software, the treatment of software in the roadmap was limited to support for hardware and networking optimizations for Big Data. We leave the more detailed discussion of the Big Data analytics applications and the data itself to the roadmap of the Big Data Value Association (BDVA).

IV. TECHNICAL DISCUSSION

A. Network

The network is the most pervasive element of any modern technology-based business. As such, the roadmap explains the potential optimizations for Big Data with innovative

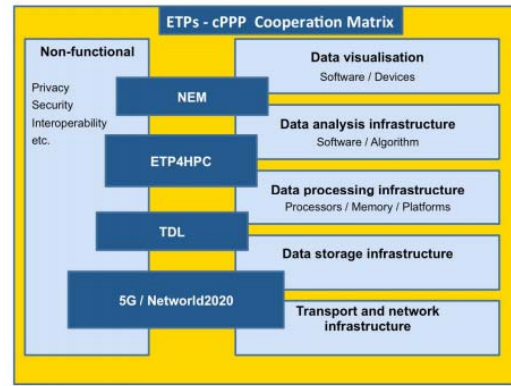


Figure 1: ETP/PPP Collaboration (from [15])

technologies applied to these appliances - specifically routers and switches - as related to virtualization.

The analysis considers network requirements for Big Data workloads, whether inside a public cloud, a private data center or even in a future High Performance / Big Data embedded system. We examine these requirements from the perspective of the “data receiving end”, meaning the network communication inside of the Data Center. As a result, we consider the nascent IoT sensors market, the Internet or mobile infrastructure challenges faced by the global telecom networks, and the actual access to the data by businesses (including regulatory and privacy concerns) from this perspective. The “network” consists of multiple functions embedded at different layers across many physical devices ranging from the server motherboard and interfaces to the top of rack switches, routers and operator infrastructure. Until now, the networking hardware lifecycle has been driven by the quest for increasing bandwidth. But today’s market landscape is rapidly changing under the pressure of demand coming from Big Data, mobile phones and IoT combined requirements.

1) Network appliance hardware: specialized to bare metal

In reaction to a competitive new landscape, hyperscalers like Google and Facebook are racing to be the first to achieve state-of-the-art bandwidth (100GE). They are also considering moving to a new architecture based on either bare metal switches or specialized “purpose-built” switches that are able to better cope with their specific Big Data workloads. Bare metal [16] refers to commodity (low-cost) switches for which customers must procure, separately, a third-party network operating system (NOS) — like Big Switch Light OS, Cumulus Linux OS, Pica8 PicOS — or build their own like Facebook did. Network operating system support and services must be obtained from the third-party NOS. Additionally, there are White Box switches that are commodity-based bare-metal switches with a preloaded network operating system from a third-party or traditional networking vendor.

2) Hardware to “softwarization” to virtualization This trend in network architecture, however, goes well beyond this bare metal hardware. The previously mentioned “softwarization” begins with Software Defined Networking (SDN) which allows for the separation of control and data planes, respectively, via software that can run on bare metal switches and or servers with the addition of network

cards [16]. This has the potential to bring down the cost significantly and can greatly increase flexibility. As explained by Google [17], SDN is about “a software control plane that abstracts and manages complexity...and can make 10,000 switches look like one.” This architecture continues with Network Function Virtualization (NFV), which allows for the implementation of security, firewalls, routing schemes and other functions separately, again via software allowing for increased control, flexibility and scalability.

3) *Deconstructing the data center (beyond 400 GbE)* High-end (beyond 400 Gigabit Ethernet or GbE) network appliances should be available after 2020 [18], but by then, the entire organization inside the data center may have changed. The continuous demand for flexibility and lower operating costs may require radical transformations [19], with high bandwidth available at all key interconnect nodes leading to composable hardware – CPU, memory, I/O and storage that is purchased à la carte and supported by software that can reconfigure the network for specific workloads.

The benefits are clear; by disaggregating the data center, we facilitate regular upgrades and potentially eliminate the need and cost of replacing entire servers, cabling and reconfiguring everything. This vision will not be realistic without new software capable of efficiently managing the complexity of such a heterogeneous pool of resources – each resource potentially located anywhere in a data center. This could lead to interesting opportunities for SMEs, due to the trend toward open hardware and networking and potentially move the ecosystem out the hands of the big vertical chip makers.

B. Architecture

Regarding Big Data compute node hardware, there are four important trends to consider, which are briefly discussed below and outlined in detail in the full roadmap [12].

1) *Heterogeneous computing* There is a noticeable trend away from general-purpose architectures towards heterogeneous systems and accelerators. This change is mainly driven by a slowdown in Moore’s Law [20][21], which leads to combinations of multiple kinds of processors and accelerators, GPUs, many-cores, FPGAs, and application-specific accelerators into the same device.

Despite the potential benefits of moving toward heterogeneous systems, the barriers to entry are substantial, in particular the cost of purchasing accelerator hardware and software complexity. The effort to run a Big Data application on heterogeneous systems requires specialized skills and knowledge of hardware due to the complex tools and programming models. Even after investing in the appropriate human capital, a suitable Return on Investment (ROI) is not guaranteed, since such systems often require hand optimization. On top of this, software for heterogeneous systems is not portable and subject to vendor lock-in. In addition, many open-source communities are philosophically opposed to accepting hardware-specific software patches [22], so only open languages and APIs are likely to be supported beyond specific driver modules connected to using general-purpose and often restrictive interfaces. Finally, many new

technologies have yet to be proven in terms of performance due to the lack of standard real-world benchmarks.

As a result, for European software vendors to adopt heterogeneous systems, they must keep pace with successive candidate technologies, which is not economically viable. This is evident in our project surveys, in which the majority of European software vendors reported that they had no hardware roadmap and preferred to wait until new technologies became widely adopted inexpensive commodities.

2) *Specialization and vendor lock-in* General-purpose GPU (GPGPU) is a maturing technology with a growing rate of adoption, especially in the area of High Performance Computing (HPC). The GPGPU market is currently dominated by Nvidia (>95% of GPU-accelerated systems in the TOP500 use Nvidia). GPGPUs have not yet achieved wide-scale penetration into data centers due the uncertain ROI. Small to medium-sized data center operators are unwilling to deploy GPGPUs at large scale, as the power consumption is too high and utilization too low to justify the investment. As is the case for moving from a GPGPU-based heterogeneous architecture to an FPGA-based one, there is considerable Non-recurring Engineering (NRE) cost required for a change in GPU vendor.

3) *Integration within the compute node* There is a trend towards greater integration, to improve performance and reduce energy consumption. Investing in a market-specific server SoC is likely to be cost-prohibitive, however, unless the design can be supported by a vertical business or it can address a large-volume market (such as mobile). SoCs provide no flexibility: adding a new interface (e.g. 40 GbE) requires a costly redesign. In addition, an SoC must be implemented using a single silicon process. Since the SoC includes the performance- and energy-critical processor cores, the die must be fabricated using an expensive leading edge silicon technology.

An alternative is System-in-Package (SiP), as pioneered by the EC EUROSERVER project [23]. Having multiple dies in the same package provides flexibility in that faster evolving technologies may be separated from more slowly evolving ones and thus replaced without affecting the rest of the design. In addition, market-specific products can be built from commodity compute chiplet(s) with specialized chiplet(s) for accelerators and I/O interfaces without designing an entire SoC. This flexibility may give smaller companies a better opportunity to compete due to tighter system integration.

4) *Verticalization and hyperscalers* The final major trend is the increasing dominance of a small number of vertically-integrated companies that co-design all or parts of the server stack, to varying degrees, ranging from the user-visible software, through (Big Data) frameworks, down to system integration and potentially even chip design. These companies have enormous market share and economies of scale, made more so through efficiencies from their vertically-integrated perspective.

In Europe, however, the industry is fragmented, with a large disconnect between technology providers and analytics companies. Almost all analytics companies expressed that they

have no hardware roadmap, take little notice of new hardware trends and are only looking at existing commodity hardware. Since Europe currently has no market share in server compute CPUs, there is limited opportunity for these companies to engage with the incumbent supplier(s). This large disconnect between technology providers and analytics companies carries a significant risk of being left behind by the larger U.S. companies.

C. Software Support

1) *Big Data processing: query languages to frameworks* In the early years of data processing, data analysts knew which answers they were looking for and their datasets were clean, so query languages were the tool-of-choice. Query languages were bound to a specific purpose, for instance SQL for relational querying, SPARQL for querying Resource Description Framework data, and XPath/XQuery for XML. Moreover, the higher-level software could be easily written in a way that was independent of the hardware.

Over the years, with the advent of Big Data, several changes in the data processing landscape have rendered query languages difficult to use. Firstly, there has been a nearly exponential increase in the volume and variety of data – data that is increasingly heterogeneous, unstructured, “dirty” and unprocessed, and therefore unsuitable for the SQL abstraction. In addition, there has been a broad shift from local to distributed computing, which has required the use of distributed frameworks, such as MapReduce, Spark and Flink, that hide the complexity of distributed hardware. These lower-level frameworks have not been directly compatible with existing query languages. The consequence has been a shift away from query languages towards data analysis libraries and APIs targeting Machine Learning (ML) and Natural Language Processing (NLP).

2) *Towards hardware dependence* The push from the business intelligence community toward Big Data analytics has resulted in widespread adoption of distributed frameworks. At the same time, open source communities are trying to provide suitable ML code higher-level libraries (MLlib) for these frameworks. Additionally, and similarly to the widespread adoption of SQL, large companies and hyperscalers have already started to develop their own solutions for NLP and ML, as specialized higher-level libraries such as IBM’s SystemT and SystemML and Google’s TensorFlow. The advantage to these companies is that these higher-level libraries can be run on nearly any distributed framework. They allow users to specify computations in a way that ensures that the program can be executed in parallel and the frameworks can then run this code on a supported set of hardware. Every time novel hardware becomes available, these frameworks need to be adapted to support the hardware.

3) *Too many abstractions* At the lower levels of the software stack, there are a large number of programming abstractions. Heterogeneous architecture approaches come with diverse programming interfaces that usually need to be addressed explicitly by Big Data application developers and

cannot be easily integrated into existing workflows automatically.

Even though every programming concept relevant for Big Data requires parallelizing work across available hardware, and even though all relevant hardware structures support parallel processing in some way, the specifics are incompatible, in that there are no common abstractions that work for everything. Abstraction for parallelism at the data center level means using MapReduce or another distributed framework while abstraction for parallelism (multicore) at node level requires yet another layer such as OpenMP (multicore), heterogeneous architectures (CPU, GPU, FPGA) require an abstraction such as OpenCL, and so on.

On a larger scale, MapReduce and its successors for batch and stream processing implemented by the Apache Spark and Apache Flink projects allow the parallel execution of code on shared-nothing clusters. All of these frameworks specify in a declarative way the data placement and unit of parallelization, while leaving the actual processing in each parallel instance to conventional functional or procedural code. The unit of parallelization supported here is an operating system thread. Any hardware that can execute such a thread, such as a CPU core, is potentially available for MapReduce, while everything else needs to be addressed explicitly by the programmer.

At the hardware level on a single node, there are similar mismatches: GPUs and CPUs are tailored for processing multiple data items at once. For each piece of hardware, however different programming approaches are necessary. GPUs rely on SIMT, and CPUs provide SIMD functionality to achieve a similar goal at least on a single CPU core. Multicore operations need to be implemented explicitly. Modern compilers and frameworks like OpenMP are capable of abstracting away some of these differences. The kernel-based programming abstractions of OpenCL translate very well to all of the previously mentioned parallelization concepts, however even OpenCL only ensure correctness of the computation on each platform. It does not ensure that the computation has been optimized for execution on these same platforms. The situation with more advanced accelerator hardware in heterogeneous architectures is similar. While FPGAs usually support standard Hardware Description Languages (HDLs) such as VHDL and Verilog, these languages describe the functionality required at a very low level, in a way that is difficult for software engineers to understand. Finally, specialized solutions such as ASICs, DSPs and neuromorphic hardware require programming interfaces that are likewise specialized and for the time being cannot be accessed automatically from conventional Big Data framework code.

V. KEY FINDINGS AND RECOMMENDATIONS

A. Industry Key Findings

Our findings are the result of 89 in-depth interviews with key stakeholders from more than 70 distinct European companies, in addition to several meetings with a broad spectrum of industry stakeholders. These companies included major and up-and-coming players from telecommunications, hardware

design and manufacturers as well as strong representation from health, automotive, financial and analytics sectors.

(1) **Industry is still focused on how to extract value from their data, and they are still looking for the business model to turn this value into profit. Consequently, they are not focused on processing (and storage) bottlenecks, let alone on the underlying hardware.** The overwhelming response is that industry does not see Big Data hardware processing problems, only Big Data value opportunities. We believe that this is largely because the industry is not yet mature enough for most companies to fully understand the kind of analytics and Big Data processing that leads to undesirable bottlenecks.

(2) **European companies are not convinced of the Return on Investment of using novel hardware.** In general, European companies are content to use currently available hardware as long as they continue to receive the most competitive pricing. In some cases this was due to extreme price-sensitivity, but for many it was simply due to risk, exacerbated by the lack of a clean metric or benchmark for side-by-side comparisons for novel hardware. Overall, the majority of the companies were not convinced that the investment in expensive hardware coupled with the person months required to make their products work with new hardware would be worthwhile.

(3) **Europe has limited opportunities for hardware / software architects to work together.** The European ecosystem is highly fragmented while media and internet giants such as Google, Amazon, Facebook, Twitter and Apple and others (also known as hyperscalers) are pursuing verticalization and designing their own infrastructures from the ground up. European companies that are not closely considering hardware and networking technologies as a means to cutting cost and offering better future services run the risk of falling further behind. Hyperscalers will continue to take risks and transform themselves because they are the “ecosystem”, moving everybody else in their trail.

(4) **Dominance of non-European companies in the server market complicates the possibility of new European entrants in the area of specialized architectures.** The vast majority of server hardware is based on Intel processors. As a result, Intel has a huge influence over the direction of the industry, and they are working to further expand this influence, including via their recent acquisition of Altera, in hopes of developing powerful new integrated hardware technologies.

B. High-level Actions Summary

The RETHINK big roadmap [12] makes the following twelve concrete recommendations. More detail is available in the complete roadmap document.

1) **Promote adoption of current and upcoming networking standards** Europe should accelerate the adoption of the current and upcoming standards (10 and 40Gb Ethernet) based on low-power consumption components proposed by European Companies and connect these companies to end users and data-center operators so that they can demonstrate their value compared to the bigger players.

2) **Prepare for the next generation of hardware and take advantage of the convergence of High Performance**

Computing (HPC) and Big Data interests In particular, Europe must take advantage of its strengths in HPC and embedded systems by encouraging dual-purpose products that bring these different communities together (e.g. HPC/Big Data hardware that can be differentiated in software). This would allow new companies to sell to a bigger market and decrease the risk associated with development of new product. There is already a clear convergence between High Performance Computing and Big Data. Moreover, large scientific experiments, including the Large Hadron Collider and Square Kilometer Array involve processing huge streams of data and are increasingly adopting Big Data technologies.

3) **Anticipate the changes in Data Center design for 400Gb Ethernet networks (and beyond)** This includes paying special attention to hardware developments such as photonics-on-silicon integration and novel Data Center interconnect designs required at 400Gb operation.

4) **Reduce risk and cost of using accelerators** Novel specialized hardware has the potential to increase computing performance and energy efficiency for appropriate applications by a factor of ten or more. In order to achieve these gains, it is necessary to re-engineer the software, which is expensive and time consuming, and it is difficult to predict the level of gains ahead of time. The use of FPGAs for computing is most prominent in financial and oil industries, with only a small number of companies addressing these markets. Europe must lower the barrier to entry for heterogeneous systems and accelerators; collaborative projects should bring together end users, application providers and technology providers to demonstrate significant (10x) increase in throughput per node on real analytics applications.

5) **Encourage system co-design for new technologies** Europe must bring together end users, application providers, system integrators and technology providers to build the most efficient integrated complete hardware—software solutions. This requires hardware to meet the evolving needs of Big Data, integrating more subsystems into the processor device as well as new non-volatile memories and I/O interfaces.

6) **Improve programmability of FPGAs** Europe should also fund research projects involving providers of tools, abstractions and high-level programming languages for FPGAs or other accelerators with the aim of demonstrating the effectiveness of this approach using real applications. Europe should also encourage a new entrant into the FPGA industry.

7) **Pioneer markets for neuromorphic computing and increase collaboration** For neuromorphic computing and other disruptive technologies, the principal issue is the lack of a market ecosystem, with insufficient appetite for risk and few European companies with the size and clout to invest in such a risky direction. Europe should encourage collaborative research projects that bring together actors across the whole chain: end users, application providers and technology providers to demonstrate real value from neuromorphic computing in real applications.

8) *Create a sustainable business environment including access to training data* Europe should address access to training data by encouraging the collection of open anonymized training data and encouraging the sharing of anonymized training data inside EC-funded projects. To address the lack of information sharing, Europe should encourage interaction between hardware providers and Big Data companies using network-of-excellence or similar.

9) *Establish standard benchmarks* It is difficult for Industry to assess the benefits of using novel hardware. We propose establishing benchmarks to compare current and novel architectures using Big Data applications.

10) *Identify and build accelerated building blocks* We propose to identify often-required functional building blocks in existing processing frameworks and to replace these blocks with (partially) hardware-accelerated implementations.

11) *Investigate use of heterogeneous resources* With edge computing and cloud computing environments calling for heterogeneous hardware platforms, we propose creation of dynamic scheduling and resource allocation strategies.

12) *Continue to ask the question – Do companies think that hardware and networking optimizations for Big Data can solve the majority of their problems?* As more companies learn how to extract value from Big Data and determine which business models lead to profits, the number of service offerings and products based on Big Data analytics will grow sharply. This growth will likely lead to an increase in consumer expectations with respect to these Big Data-driven products and services, and we expect companies to run into more and more undesirable performance bottlenecks that will require optimized hardware.

VI. CONCLUSIONS

This paper describes the RETHINK big Project and its methodology, and it gives a brief overview of the key findings and recommendations in the RETHINK big roadmap [12]. In summary, the RETHINK big Project has created an industry-led strategic roadmap that will maximize European industry competitiveness for Big Data hardware over the next 10 years.

ACKNOWLEDGMENT

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 619788. It has also been supported by the Spanish Government (grant SEV2015-0493 of the Severo Ochoa Program), by the Spanish Ministry of Science and Innovation (contract TIN2015-65316) and by Generalitat de Catalunya (contracts 2014-SGR-1051 and 2014-SGR-1272).

REFERENCES

- [1] G. Press, "12 Big Data Definitions: What's Yours?" 2014 [Online], Available: <http://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#32c2a68821a9>.
- [2] G. Nunberg, "Forget YOLO: Why 'Big Data' Should Be The Word Of The Year," 2012 [Online], Available: <http://www.npr.org/2012/12/20/167702665/geoffrey-nunbergs-word-of-the-year-big-data?live=1>.
- [3] Helpdesk response, "What is Big Data and what classifies as Big Data?" 2016 [Online], Available: <http://stackoverflow.com/questions/35560823/what-is-big-data-what-classifies-as-big-data>.
- [4] A. Putnam, et al. "A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services," Proceedings of the 41st Annual International Symposium on Computer Architecture (ISCA 2014), 2014 [Online], Available: <https://www.microsoft.com/en-us/research/publication/a-reconfigurable-fabric-for-accelerating-large-scale-datacenter-services/>.
- [5] T. Prickett Morgan, "Microsoft Extends FPGA Reach from Bing to Deep Learning," 2015 [Online], Available: <http://www.nextplatform.com/2015/08/27/microsoft-extends-fpga-reach-from-bing-to-deep-learning/>.
- [6] M. Feldman, "Microsoft Goes All in for FPGAs to Build Out AI Cloud," 2016 [Online], Available: <https://www.top500.org/news/microsoft-goes-all-in-for-fpgas-to-build-out-cloud-based-ai/>.
- [7] A. Chanthadavong, "Nvidia Touts GPU Processing as the Future of Big Data," 2016 [Online], Available: <http://www.zdnet.com/article/nvidia-touts-gpu-processing-as-the-future-of-big-data/>.
- [8] A. Woodie, "GPU-Powered Deep Learning Emerges to Carry Big Data Torch Forward," 2016 [Online], Available: <http://www.datanami.com/2016/04/13/gpu-powered-deep-learning-emerges-carry-big-data-torch-forward/>.
- [9] DRIVE PX, Nvidia website, <http://www.nvidia.com/object/drive-px.html> (accessed 25 Nov 2015).
- [10] N. Ahmad in interview with T. Prickett Morgan, "Facebook Ops Director On Breaking Open The Switch" 2015 [Online], Available: <http://www.nextplatform.com/2015/03/16/facebook-ops-director-on-breaking-open-the-switch/>.
- [11] The Open Compute Initiative, <http://www.opencompute.org/>.
- [12] RETHINK big D5.3 Roadmap [Online], Available: <http://www.rethinkbig-project.eu/roadmap>.
- [13] Earl Joseph, IDC, "HPC Trends, Big Data and the Emerging Market for High Performance Data Analysis (HPDA)" 2013 [Online], Available: <http://www.lanl.gov/conferences/salishan/salishan2013/IDC%20Salishan%20Presentation%204.18.2013.pdf> (for BDEC).
- [14] H. Alkhatib, P. Faraboschi, E. Frachtenberg, et. al. IEEE CS 2022 Report, Feb 2014 [Online], Available: <https://www.computer.org/cms/ComputingNow/2022Report.pdf>.
- [15] BDVA SRIA v2.0, Jan 2016 [Online], Available: http://www.bdva.eu/sites/default/files/EuropeanBigDataValuePartnership_SRIA_v2.pdf.
- [16] A. Kindness – Forrester blog, "What is the difference between white box bare metal branded bare metal and oem network switches?" 2015 [Online], Available: http://blogs.forrester.com/andre_kindness/15-02-23-what-is-the-difference-between-white-box-bare-metal-branded-bare-metal-and-oem-network-switches.
- [17] T. Hoff - High Scalability blog, "How Google Invented An Amazing Datacenter Network Only They Could Create" 2015 [Online], Available: <http://highscalability.com/blog/2015/8/10/how-google-invented-an-amazing-datacenter-network-only-they.html>
- [18] D. Coleman, "Optical Trends in the Data Center" 2012 [Online], Available: https://www.bicsi.org/uploadedFiles/BICSI_Website/Global_Community/Presentations/Caribbean/1.07%20Corning.pdf.
- [19] T. Prickett Morgan, "Smashing the Server to Put it Back Together Again" 2015 [Online], Available: <http://www.nextplatform.com/2015/11/04/smashing-the-server-to-put-it-back-together-better/>.
- [20] A. Huang, "The Death of Moore's Law Will Spur Innovation," IEEE Spectrum March 2015 [Online], Available: <http://spectrum.ieee.org/semiconductors/design/the-death-of-moores-law-will-spur-innovation>.
- [21] M. Mitchell Waldrop, "The chips are down for Moore's law", Nature, Vol. 520, Issue 7589. [Online], Available: <http://www.nature.com/news/the-chips-are-down-for-moores-law-1.19338>.
- [22] LWN.net, "Linux and TCP offload engines," 2005 [Online] Available: <http://lwn.net/Articles/148697>.
- [23] M. Marazakis et al. EUROSERVER: Share-Anything Scale-Out Micro-Server Design. Design, Automation and Test in Europe (DATE), 2016