BeeGFS in the DEEP/-ER Project

Cristina Manzano, Jülich Supercomputing Centre

DEEP and DEEP-ER

www.deep-project.eu www.deep-er.eu

EU-Exascale projects
20 partners
Total budget: 28,3 M€
EU-funding: 14,5 M€
Nov 2011 – Sept 2016

Visit us @
ISC'16, Frankfurt (Germany)
20.-22.06.2016

-Booth
-BoF
-Workshop

What are the projects about?

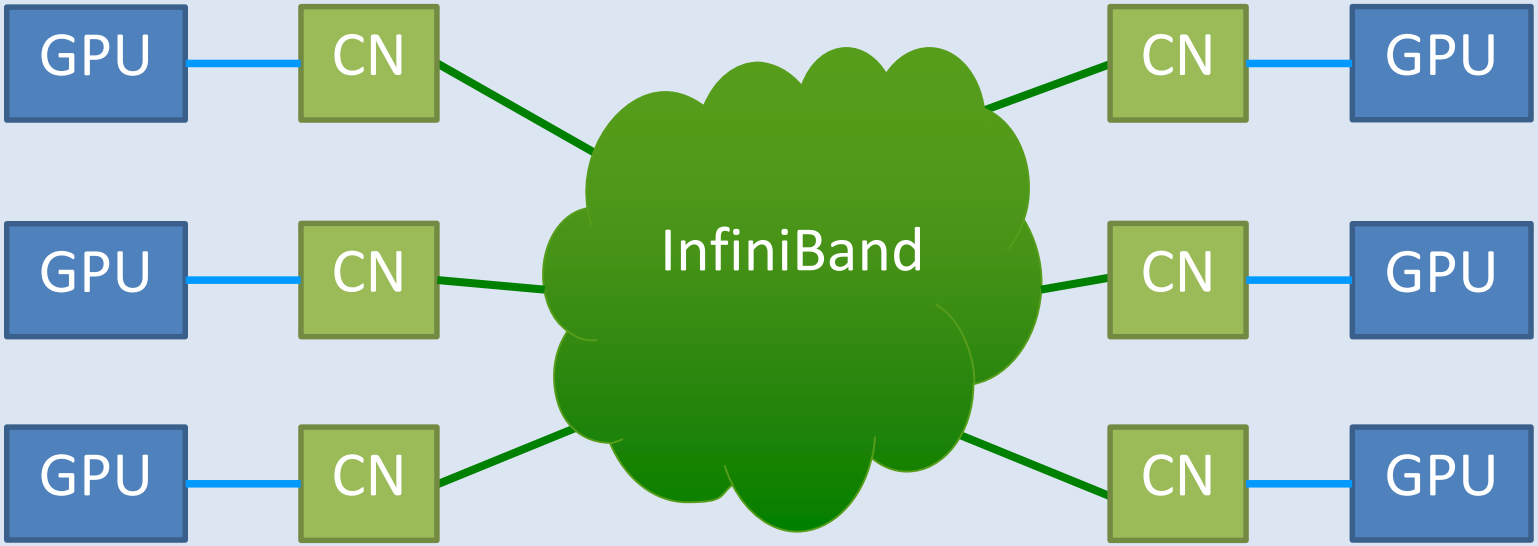
DEEP

- **Cluster-Booster archit.**
- Software stack
- Programming environ.
- Energy efficiency
- Applications:
  - Co-design
  - Evaluation/demonstration
  - Code modernisation

DEEP-ER

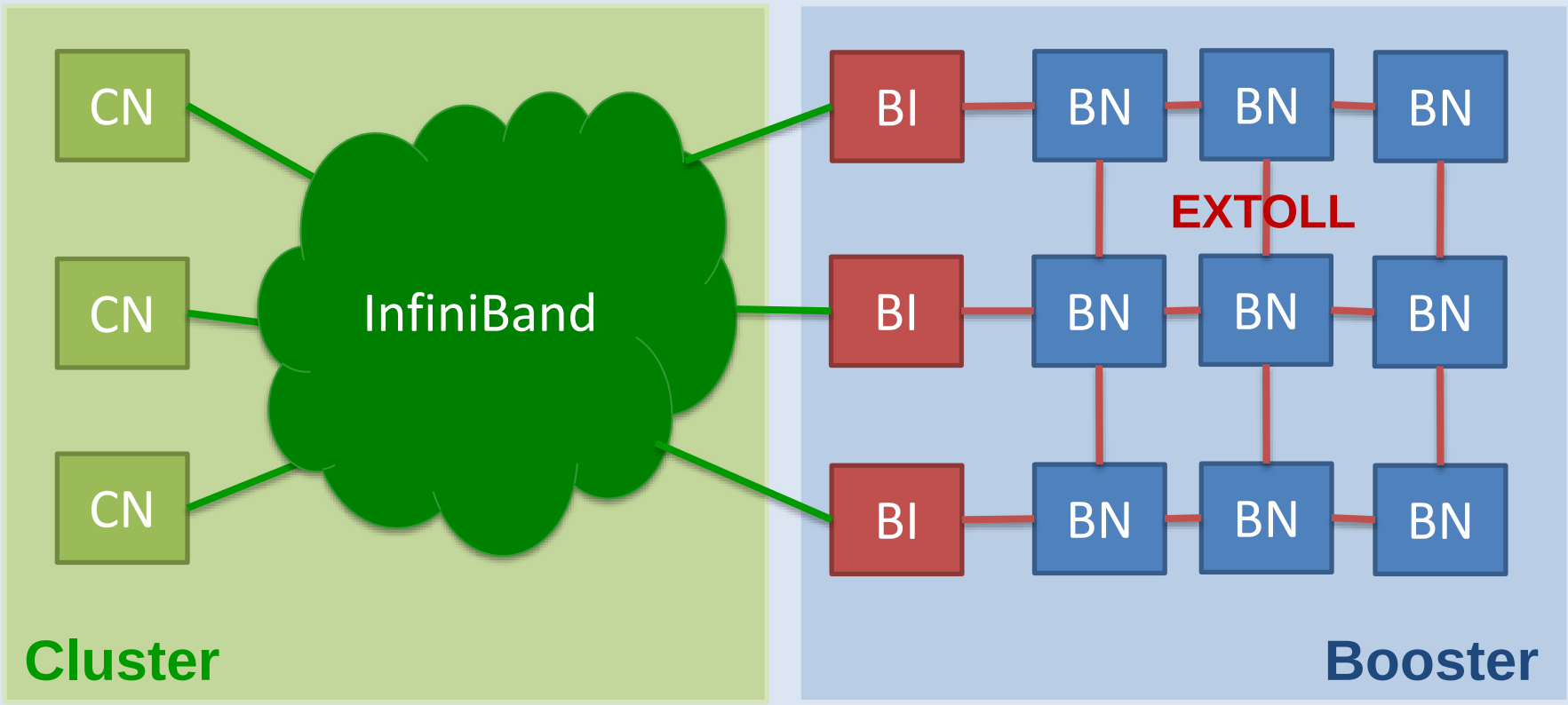
- Extend memory hierarchy
- High-performance **I/O**
- Scalable **resiliency**
- Applications:
  - Co-design
  - Evaluation/demonstration
  - Code modernisation

“Standard” heterogeneity



- Flat topology
- Simple management of resources
- Static assignment of accelerators to CPUs
- Accelerators cannot act autonomously

DEEP-ER Cluster-Booster architecture



Flexible assignment of resources (CPUs, accelerators)
Direct communication between accelerators
“Offload” of large and complex parts of applications

DEEP System

- Installed at JSC
- 1,5 racks
- 500 TFlop/s peak perf.
- 3.5 GFlop/s/W
- Water cooled

Cluster (128 Xeon Sandy Bridge)

Booster (384 Xeon Phi KNC)

DEEP System

File Servers (6 Xeon Sandy Bridge)

JBOD 2245 (45x 2TB disks)

- /work file system
- ~2000 MB/s write/read BW*

* Measured with IOR benchmark.

BeeGFS @ DEEP

DEEP Storage servers

- 6x DELL PowerEdge R520 storage servers (deep-fs01 – deep-fs06)

SAS switch

- 1x LSI 6140 SAS switch connecting the storage servers with the JBOD

JBOD

- 1x SGI JBOD 2245 with 45x 2TB disks

Storage space on each server

deep-fs01	RAID1: 2x mirrored disks
deep-fs02	RAID1: 2x mirrored disks
deep-fs03	RAID6: 10x disks
deep-fs04	RAID6: 10x disks
deep-fs05	RAID6: 10x disks
deep-fs06	RAID6: 10x disks

BeeGFS @ DEEP

BeeGFS configuration

Node	Description	BeeGFS roles	BeeGFS services
deep-fs01	Storage node	Management, Metadata, Administration, Monitoring, Helper	beegfs-mgmtd, beegfs-meta, beegfs-admon, beegfs-helperd
deep-fs02	Storage node	Metadata, Helper	beegfs-meta, beegfs-helperd
deep-fs0[3-6]	Storage nodes	Storage, Helper	beegfs-storage, beegfs-helperd
deep[1-128]	Compute nodes	Client, Helper	beegfs-client, beegfs-helperd
deepm	Administration (master) node	Client, Helper	beegfs-client, beegfs-helperd
deepl	Login node	Client, Helper	beegfs-client, beegfs-helperd

What are the projects about?

DEEP

- **Cluster-Booster archit.**
- Software stack
- Programming environ.
- Energy efficiency
- Applications:
  - Co-design
  - Evaluation/demonstration
  - Code modernisation

DEEP-ER

- Extend memory hierarchy
- High-performance **I/O**
- Scalable **resiliency**
- Applications:
  - Co-design
  - Evaluation/demonstration
  - Code modernisation

What are the projects about?

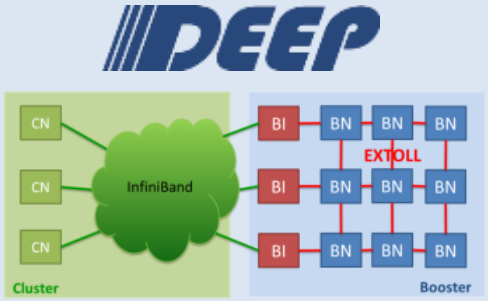
DEEP

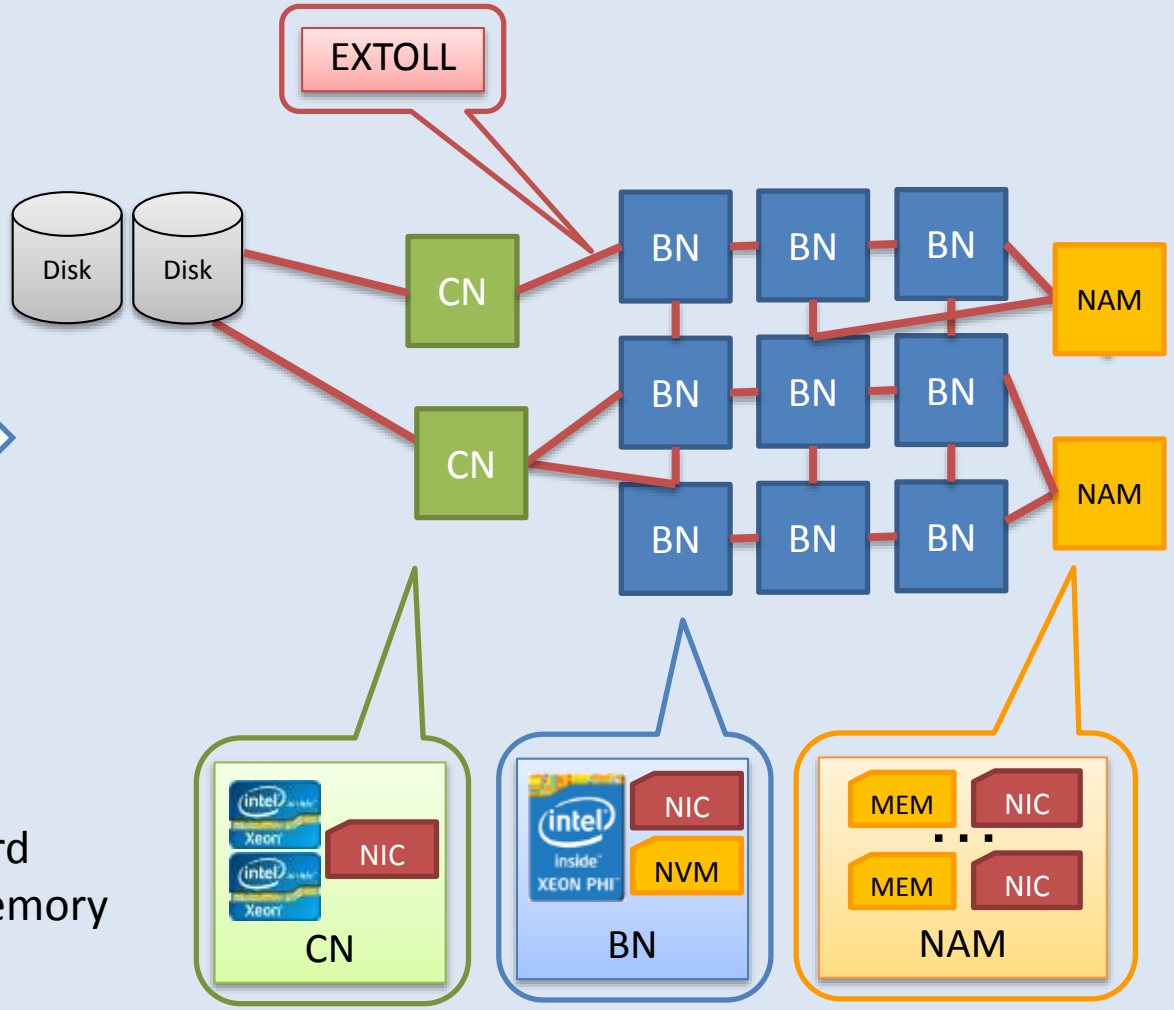
- **Cluster-Booster archit.**
- Software stack
- Programming environ.
- Energy efficiency
- Applications:
  - Co-design
  - Evaluation/demonstration
  - Code modernisation

DEEP-ER

- Extend memory hierarchy
- High-performance **I/O**
- Scalable **resiliency**
- Applications:
  - Co-design
  - Evaluation/demonstration
  - Code modernisation

Enhance DEEP architecture



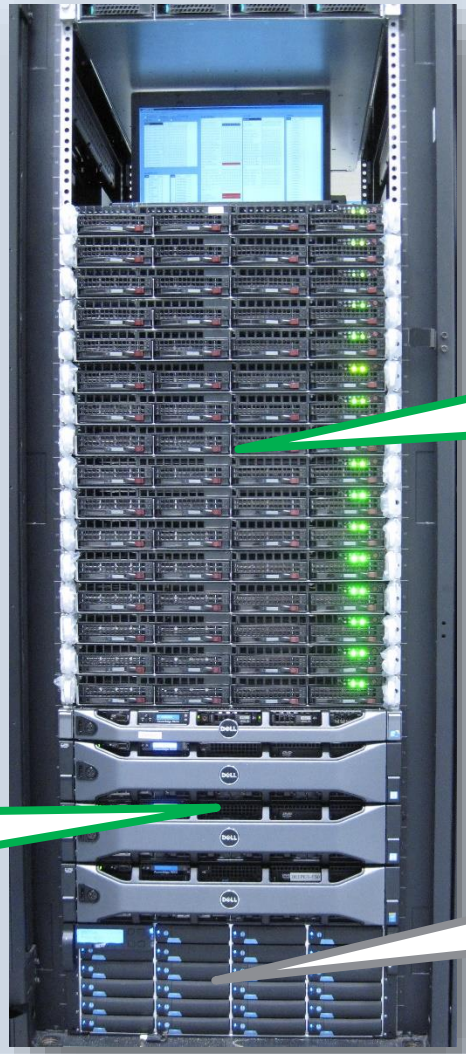


Legend:

CN:	Cluster Node
BN:	Booster Node
NIC:	Network Interface Card
NAM:	Network Attached Memory
NVM:	Non Volatile Memory

Software Development Vehicle (SDV)

- /sdv-work file system
- ~1500 MB/s write/read BW*



Cluster (16 Xeon Haswell)

File Servers (3 Xeon Haswell)

RAID EUROstor (24x 6TB disks)

* Measured with BeeGFS benchmark. IOR benchmarking ongoing work.

BeeGFS @ DEEP-ER

DEEP-ER Storage servers

- 3x DELL PowerEdge R530 storage servers (deeper-fs01 – deeper-fs03)

Metadata

- 2x internal SSD disks

RAID System

- 1x EUROstor ES-6600 with 4 x 8Gbit FC connector
- 24x 6 TB SAS Nearline (RAID6)
- 4x 31500.0GB Volumes (2 unused for future expansion of the storage system)

Storage space on each server

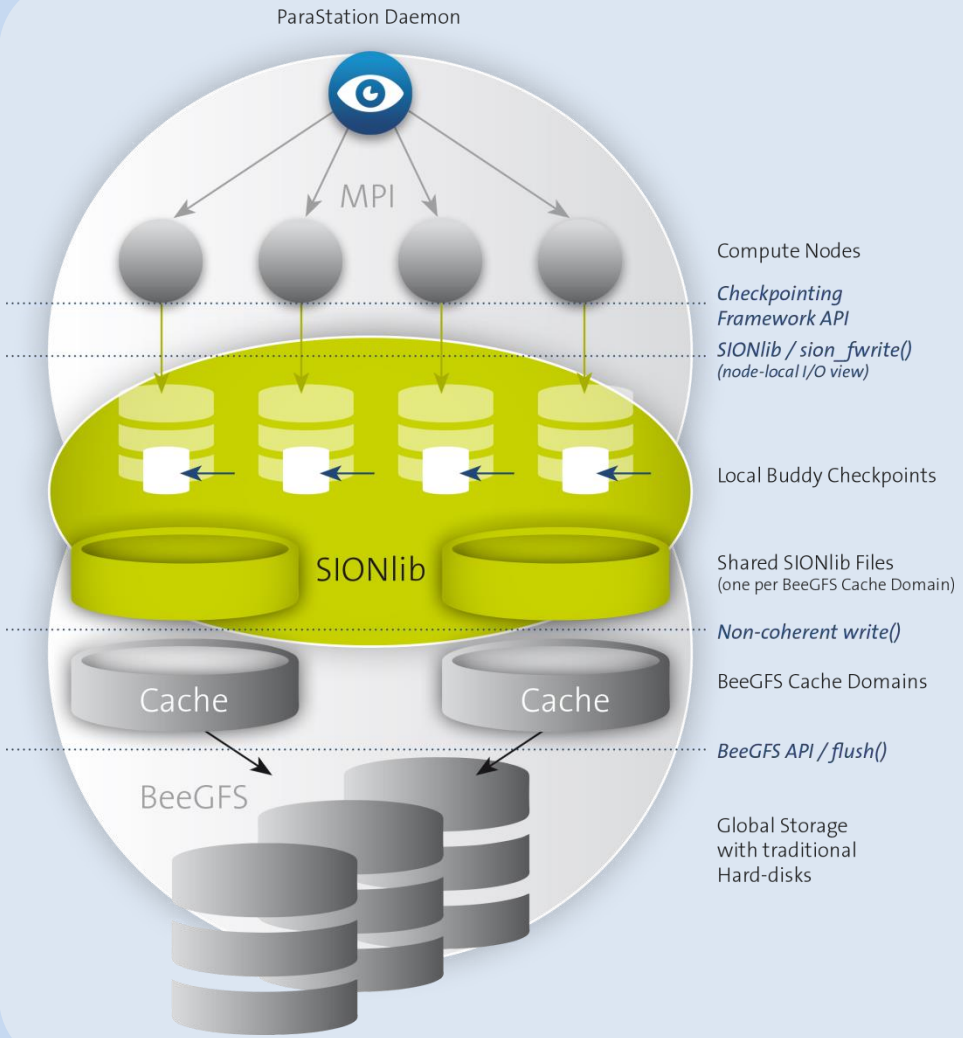
deeper-fs01	RAID1: 2x internal SSD disks (mirrored)
deeper-fs02	1x 31500.0GB Volume
deeper-fs03	1x 31500.0GB Volume

BeeGFS @ DEEP-ER

BeeGFS configuration

Node	Description	BeeGFS roles	BeeGFS services
deeper-fs01	Storage node	Management, Metadata, Administration, Monitoring, Helper	beegfs-mgmtd, beegfs-meta, beegfs-admon, beegfs-helperd
deeper-fs0[2-3]	Storage node	Storage, Helper	beegfs-storage, beegfs-helperd
deeper-sdv[01-16]	Compute nodes	Client, Helper	beegfs-client, beegfs-helperd
deepm	Administration (master) node	Client, Helper	beegfs-client, beegfs-helperd
deepl	Login node	Client, Helper	beegfs-client, beegfs-helperd

I/O and Resiliency in DEEP-ER



Optimized I/O

- Hierarchical global FS
  - Fast caches (NVMe)
- SIONlib & E10
  - Address the “small I/O” problem

Enhanced resiliency

- Enhanced SCR
  - Built on top of the optimized I/O
- Task-based resiliency

BeeOND @ DEEP-ER

NVMe SSD devices

- NVM component: Intel DC P3700
  - -> 20nm MLC NAND Flash technology
  - -> PCI Express generation 3 × 4
- 1 NVMe with 400 GB attached to each node in Cluster and Booster
- 1 BeeOND instance running on each NVMe device
- BeeGFS cache layer
  - -> Local tier in a multi-tier storage environment
  - -> Burst buffer for temporary storage (like checkpointing)
- More about this in Frank's talk later today!

BeeOND @ DEEP-ER

Performance NVMe ext4 vs. BeeOND running on NVMe:

BeeOND @ DEEP-ER

testdir	itemspertask	filesperdir	FCreateMax [ops/sec]	FRemoveMax [ops/sec]
/nvme/tmp/	41666	651	127.947,63	58.818,83
/mnt/beeond/	41666	651	12.158,12	16.653,14

BeeOND @ DEEP-ER

testdir	API	Access	blockSize	transferSize	Aggregatesize [GiB]	Wrbwmax [MiB/s]	Rdbwmax [MiB/s]
/nvme/tmp/	POSIX	file-per-process	10GB	16MB	240 GiB	1026,68	2174,92
/mnt/beeond/	POSIX	file-per-process	10GB	16MB	240 GiB	979,76	2347,14
/nvme/tmp/	MPIIO	file-per-process	10GB	16MB	240 GiB	1121,32	1755,51
/mnt/beeond/	MPIIO	file-per-process	10GB	16MB	240 GiB	1118,76	1797,49
/nvme/tmp/	POSIX	single-shared-file	10GB	16MB	240 GiB	816,07	3001,10
/mnt/beeond/	POSIX	single-shared-file	10GB	16MB	240 GiB	406,95	2168,83
/nvme/tmp/	MPIIO	single-shared-file	10GB	16MB	240 GiB	842,01	2490,83
/mnt/beeond/	MPIIO	single-shared-file	10GB	16MB	240 GiB	425,52	2039,94

Some lessons learned

- BeeGFS is really easy to update also between major releases
  -> Script provided for updating between 2014 (FhGFS) and 2015 (BeeGFS)
- BeeGFS runs really stable
- Don’t underestimate the use of extended attributes!
  -> Gain factor of 50 (from 130 to 6300 files/second with an mdtest)
- Some users want to be able to change the stripping settings
  -> New feature in a future BeeGFS release?
- Managing BeeOND instances: clean cache after each job, start/stop services, …
  -> Developing scripts and integrating them in the ParaStation cluster management tools
- BeeGFS Client on Xeon Phi
  -> We need to provide access to the work file system also on the Booster nodes

Future challenges

- Improve performance of BeeGFS over EXTOLL

  -> Our colleagues in Fraunhofer already working in developing native EXTOLL support

- BeeOND on the NVMes in the Booster

  -> Besides the client, other services need to be installed and configured on the Xeon Phi: beegfs-mgmtd, beegfs-meta, beegfs-storage, …

BACKUP

Interested?

Contact us!

DEEP
pmt@deep-project.eu

DEEP-ER
pmt@deep-er.eu

LinkedIN
http://linkd.in/1KiBe3y

Twitter
@DEEPprojects

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n° 287530 and n° 610476

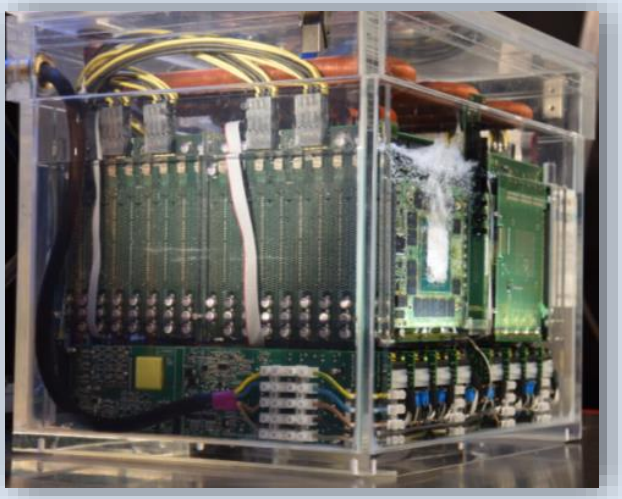
GreenICE system

Alternative Booster implementation

- Interconnect EXTOLL ASIC “Tourmalet”
- 32 KNC-node system
- Implement 4×4×2 topology, with Z dimension open

Experiment 2-phase immersion cooling

- NOVEC liquid from 3M
- Evaporates at about 50 degrees
- Condensates again in a water cooling pipe
- Allows very high-density integration



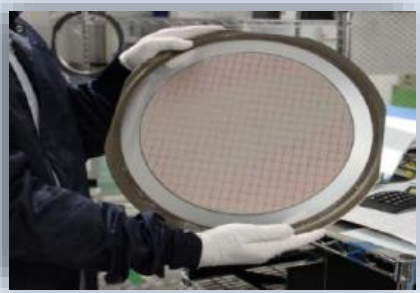
GreenICE Booster

Interconnect

	EXTOLL	Intel True Scale		Mellanox IBAN		PLX Technology
	Tourmalet	QDR	QDR 80	EDR	FDR	ExpressFabric®
Availability	Q3/2015	Now	Now	2015	Now	2015
Switches	None	IBAN	IBAN	IBAN	IBAN	PCIe switches req.
Topologies	≤7 direct connections	Switched, any, 1 rail	Switched, any 2 rails	Switched, any, 1-2 rails	Switched, any, 1-2 rails	Switched, any, 1 rail only
# Links per NIC	7	1 or 2	1 or 2	1 or 2	1 or 2	1-4 (for DEEP-ER)
Link BW	120 Gbit/s	40 Gbit/s	80 Gbit/s	103 Gbit/s	56 Gbit/s	32 (4 links) –128 (1 link) Gbit/s
Aggregate BW	940 Gbit/s	80 Gbit/s	160 Gbit/s	206 Gbit/s	112 Gbit/s	128 Gbit/s
# contexts	256	64	2*64			64
SR-IOV support	No	No	No	No	Yes	Yes
Drivers & Firmware	Adaptable	Available	Available	N/A	Available, KNL?	OSS
Driver I/F	VELO, SMFU, OFED	OFED, PSM	ODEF, PSM	OFED	OFED	OFED

EXTOLL Tourmalet

Tourmalet PCI Express Board



Tourmalet Chip and Wafer

TOURMALET

Main EXTOLL characteristics

- Direct network: no switches required
- Integrates network interface controller
- Supports 6+1 links
- Capable of tunneling PCIe (allows remote-booting KNC from the network)

Current version of EXTOLL ASIC

- 270 million transistors
- Link bandwidth: 100 G
- MPI latency: 850 ns
- MPI bandwidth: 8.5 GB/s
- Message rate: 70 million mgs/sec
- PCIe Gen3 x16