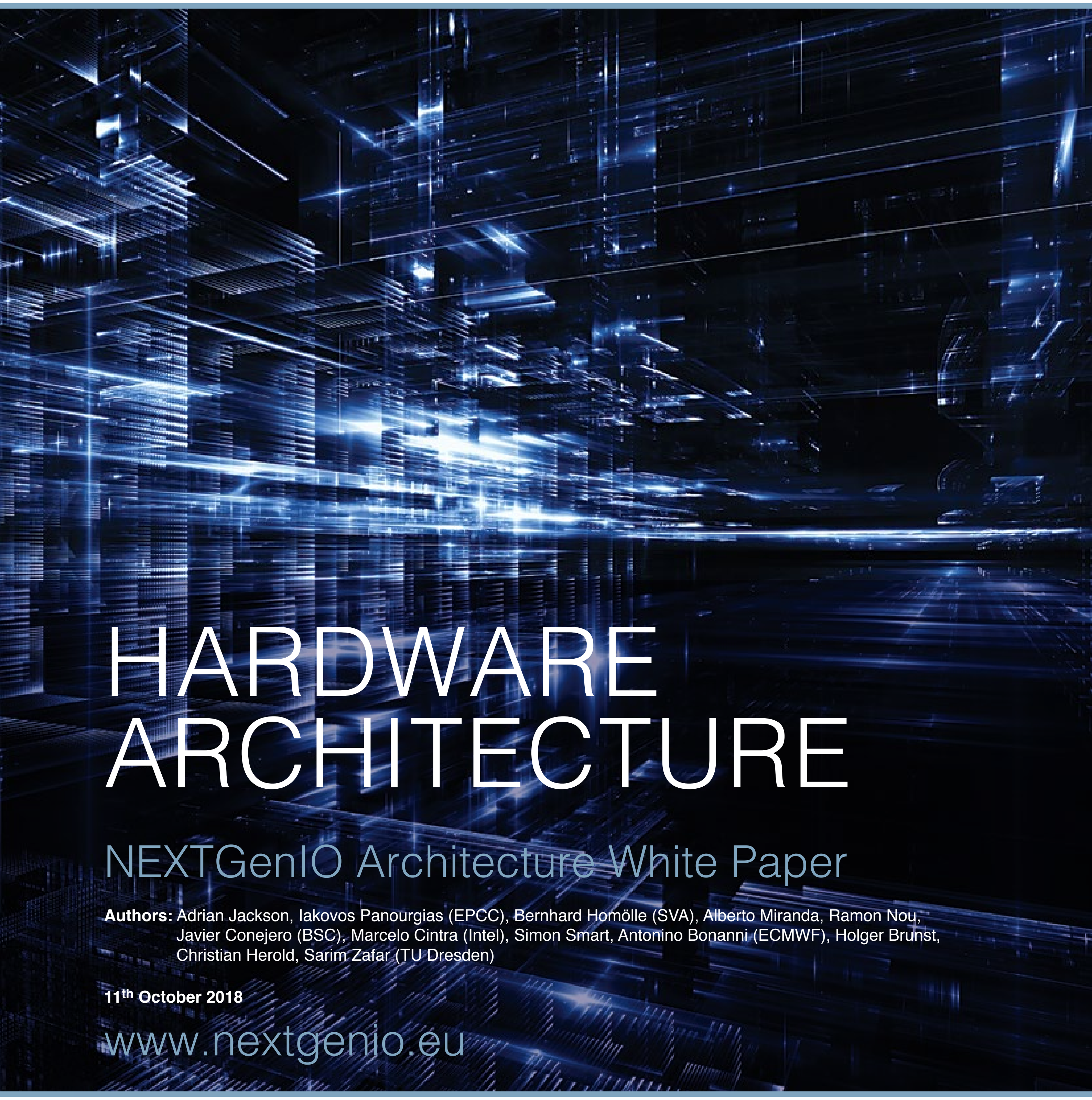

HARDWARE ARCHITECTURE

NEXTGenIO Architecture White Paper

Authors: Adrian Jackson, Iakovos Panourgias (EPCC), Bernhard Homölle (SVA), Alberto Miranda, Ramon Nou, Javier Conejero (BSC), Marcelo Cintra (Intel), Simon Smart, Antonino Bonanni (ECMWF), Holger Brunst, Christian Herold, Sarim Zafar (TU Dresden)

11th October 2018

www.nextgenio.eu

Next Generation I/O for the Exascale
This project has received funding from the European Union's Horizon 2020
Research and Innovation programme under Grant Agreement no. 671951

Contents

Table of Contents

Table of Figures

Table of Tables

Foreword

BY DR MICHÈLE WEILAND, NEXTGENIO PROJECT MANAGER

NEXTGenIO is working on improved I/O performance for HPC and data analytics workloads. The project is building a prototype hardware system with byte-addressable persistent memory on the compute nodes, as well as developing the systemware stack that will enable the transparent use of this memory.

Another key outcome from the project's research is the publication of a series of White Papers, covering topics ranging from the architecture requirements and design, to the systemware layer and applications. Each White Paper addresses one of the core challenges or developments that were addressed as part of NEXTGenIO.

This White Paper describes the hardware architecture that was developed based on a detailed requirements capture exercise and that is implemented in the NEXTGenIO prototype.

About nextgenio

Current HPC systems perform on the order of tens to hundreds of petaFLOPs. Although this already represents one million billion computations per second, more complex demands on scientific modelling and simulation mean even faster computation is necessary. The next step is Exascale computing, which is up to 1000x faster than current Petascale systems. Researchers in HPC are aiming to build an HPC system capable of Exascale computation by 2022.

One of the major roadblocks to achieving this goal is the I/O bottleneck. Current systems are capable of processing data quickly, but speeds are limited by how fast the system is able to read and write data. This represents a significant loss of time and energy in the system. Being able to widen, and ultimately eliminate, this bottleneck would significantly increase the performance and efficiency of HPC systems.

NEXTGenIO will solve the problem by bridging the gap between memory and storage. This will use Intel's revolutionary new Optane DC Persistent Memory, which will sit between conventional memory and disk storage. NEXTGenIO will design the hardware and software to exploit the new memory technology. The goal is to build a system with 100x faster I/O than current HPC systems, a significant step towards Exascale computation.

The advances that Optane DC Persistent Memory and NEXTGenIO represent are transformational across the computing sector.

1 Executive Summary

NEXTGenIO is developing a prototype high-performance computing (HPC) and high-performance data analytics (HPDA) system that integrates byte-addressable storage class memory (SCM) into a standard compute cluster to provide greatly increased I/O performance for computational simulation and data analytics tasks.

To enable us to develop a prototype that can be used by a wide range of computational simulation application, and data analytic tasks, we have undertaken a requirements-driven design process to create hardware and software architectures for the system. These architectures both outline the components and integration of the prototype system, and define our vision of what is required to integrate and exploit SCM to enable a generation of Exascale systems with sufficient I/O performance to ensure a wide range of workloads can be supported.

The hardware architecture, which is the focus of this White Paper, is designed to scale up to an ExaFLOP system. It uses high-performance processors coupled with SCM in NVRAM (non-volatile random access memory) form, traditional DRAM memory, and an Omni-Path high-performance network, to provide a set of complete compute nodes that can undertake both HPC and HPDA workloads.

2 Introduction

The NEXTGenIO project is developing a prototype HPC system that utilises byte-addressable SCM hardware to provide greatly improved I/O performance for HPC and HPDA applications. A key part of developing the prototype is the design of the different components of the system.

This document outlines the architectures we have designed for the NEXTGenIO prototype and future Exascale HPC/HPDA systems, building on a detailed requirements capture undertaken in the project. Our aim in this process has been to design a prototype system that can be used to evaluate and exploit SCM for large scale computations and data analysis, and to design a hardware and software architecture that could be exploited to integrate SCM with Exascale-sized HPC and HPDA systems.

NEXTGenIO is building a hardware prototype using Intel's Optane™ DC Persistent memory (DCPMM), which is based on 3D XPoint™ technology, and we have designed a software architecture that is applicable to other instances of SCM once they become available. The functionality outlined is sufficiently generic that it can exploit a range of different hardware to provide the persistent and high-performance storage functionality that DCPMM offers.

The remainder of this document outlines the key features of SCM that we are exploiting in NEXTGenIO, the general requirements that we have identified and the hardware architecture that we have designed.

2.1 Glossary of Acronyms

Acronym	**Description**
1GE	1 Gigabit Ethernet
10GE	10 Gigabit Ethernet
100G	100 Gigabit (dual simplex)
3D XPoint™	Intel's emerging SCM technology that will be available in NVMe SSD and NVDIMM forms
AEP	Apache Pass, code name for Intel NVDIMMs based on 3D XPointTM memory
CPU	Compute processing unit
CCN	Complete compute node
DCPMM	Data Centre Persistent Main Memory
DDR4	Double data rate DRAM access protocol (version 4)
DIMM	Dual inline memory, the mainstream pluggable form factor for DRAM
DMA	Direct memory access (using a hardware state machine to copy data)
DP	Dual processor
DPA	DIMM physical address
DP Xeon	DP Intel® Xeon® platform
DRAM	Dynamic random access memory
ExaFLOP	1018 FLOP/s
FLOP	Floating point operation (in our context 64bit/dual precision)
Gb, GB	Gigabyte (10243 bytes, JEDEC standard)
GIOPS	109 IOPS
HDD	Hard disk drive. Primarily used to refer to spinning disks
HPC	High Performance Computing
IB	InfiniBand
IB-EDR	IB "eight data rate" (100Gb/s per link)
IFT card	Internal faceplate transition card (interfaces internal iOPA cables coming from CPUs to server faceplate)
InfiniBand	Industry standard HPC communication network
I/O	Input / output

iOPA	Integrated OPA (with HFI on the CPU package, as opposed to having it on a PCIe card)
IOPS	I/O operations per second
IPMI	Intelligent Platform Management Interface
JEDEC	JEDEC Solid State Technology Association, formerly known as the Joint Electron Device Engineering Council; governs DRAM specifications
Lustre	Parallel distributed file system; Open source
LNET, LNET	Lustre network protocol
MIOPS	106 IOPS
MPI	Message Passing Interface, standard for inter-process communication heavily used in parallel computing
MW	Mega Watt (106 Watts)
NAND flash	Mainstream NVM used in SSDs today
NUMA	Non-uniform memory access
NVM	Non-volatile memory (generic category, used in SSDs and SCM)
NVMe	NVM Express, a transport interface for PCIe attached SSDs
NVMF	NVM over fabrics, the remote version of NVMe
NVRAM	Non-volatile RAM (DIMM based). Implemented by NVDIMMs, being one kind of NVM, that supports both non-volatile RAM access and persistent storage
NVDIMM	Non-volatile memory in DIMM form fact that supports both non-volatile RAM access and persistent storage. Intel NVDIMM, also known as AEP.
OEM	Original equipment manufacturer
Omni-Path	High-performance interconnect fabric developed by Intel for use in HPC systems
OPA	Omni-Path
O/S	Operating system
PB	Petabyte (10245 bytes, JEDEC standard)
PCI-Express	Peripheral Component Interconnect Express. Communication bus between processors and expansion cards or peripheral devices.
PCIe	PCI-Express
PERSISTENT	Data that is retained beyond power failure
PFLOP	Peta-FLOP, 1015 FLOP/s
PMEM, pmem	Persistent memory, another name for SCM
PXE	Pre-boot Execution Environment, run before O/S is booted
QPI	Quick Path Interconnect
Quick Path Interconnect	Intel term for the communication link between Intel® Xeon® CPUs on a DP system
RAM	Random access memory
RDMA	Remote DMA, DMA over a network
SCM	Byte-Addressable Storage class memory, referring to memory technologies that can bridge the huge performance gap between DRAM and HDDs while providing large data capacities and persistence of data
SNIA	Storage Networking Industry Association
SNMP	Simple Network Management Protocol
SSD	Solid state disk drive
SHDD	Solid state HDD, using an NVM buffer inside a HDD
TB	Terabyte (10244 bytes, JEDEC standard)
TFTP	Trivial File Transfer Protocol
U	Rack height unit (1.75”)
Xeon®	Intel’s brand name for x86 server processors
Xeon Phi™	Intel’s manycore x86 line of CPUs for HPC systems

3 Hardware Architecture

Developing a prototype platform to test and evaluate the performance and functionality benefits of SCM is one of the key objectives of the NEXTGenIO project. The SCM technology we are using, 3D XPoint™ memory, is a new type of memory that requires specific hardware support. 3D XPoint™ memory cannot simply be installed in an existing system.

However, one of the major benefits of SCM (or NVRAM) is that it comes in the same hardware form factor as standard memory (DRAM, DDR4 for 3D XPoint™). This means we can design and build a system that can take both standard memory and SCM in the same node and provide both to users.

SCM does required specific processor support, primarily because memory controllers in modern computing systems are directly integrated into the processor itself. This means, if we want to use a new memory technology we need processors with support for that new memory in their memory controllers. For instance, for DCPMM, Intel will release a new generation of Intel® Xeon® processors with the code name "Cascade Lake" that will include such support.

Furthermore, as we are combining two types of memory in the servers (or nodes) in the system, we need to ensure that there is capacity to install sufficient amounts of standard memory and SCM. As we are looking to provide a prototype system where large-scale computational simulation and data analytics tasks can be undertaken, we need to provide sufficient hardware capacity to install sufficiently large amounts of standard DRAM memory and SCM at the same time. In other words, we need a system that has space for a large number of memory modules, or DIMMs, to be installed.

The NEXTGenIO prototype system is designed to provides for very high SCM capacities and I/O bandwidth, combining is 3D XPoint™ technology with Intel's new Omni-Path interconnect to provide extremely high I/O performance for applications.

A single NEXTGenIO rack holds up to 32 Xeon® servers, Complete Compute Nodes (CCNs), connected by the Omni-Path network, with two login nodes to support local compilation, visualisation, and rack management. Furthermore, two boot servers are provided to store and serve the operating system and software required by the CCNs, and two gateway servers to enable access to an external filesystem for long term storage, and for performance comparison with the I/O provided by the SCM.

The architecture is designed with scalability in mind, facilitating the connection of NEXTGenIO racks into larger systems, reaching out well into the ExaFLOP range. We enable software to access local SCM almost as fast as DRAM, and to access and any other CCN's SCM in the cluster faster than accessing a fast storage device (i.e. a fast SSD) inside the node itself.

This SCM focussed system architecture allows fundamentally new designs for scalable HPC software and provides the functionality required to make the data centre power consumption become the final scaling limiter, not the system architecture.

Using future processors and network generations, we expect an ExaFLOP configuration based on the NEXTGenIO architecture to become practical within five years.

3.1 Hardware building blocks

System or application software can use several different NVDIMM libraries, such as libmemkind or PMDK. Echofs is using PMDK libraries to provide the needed persistency for the write operations to the NVRAM space.

3.1.1 Main server node

Since NEXTGenIO is centred on using SCM as a new, ultra-fast storage layer between memory (DRAM) and disk, we have to understand the goals and constraints on how to configure DRAM and NVM in the servers being used.

As shown in Figure 1, the NEXTGenIO server node is a dual processor node and supports SCM. As we are looking for high performance nodes, we need to have all 12 memory channels (as provided by Intel's current "Skylake" generation of Intel® Xeon® processors and announced for "Cascade Lake") populated with some DRAM, to enable utilising the total available DRAM bandwidth in the node.

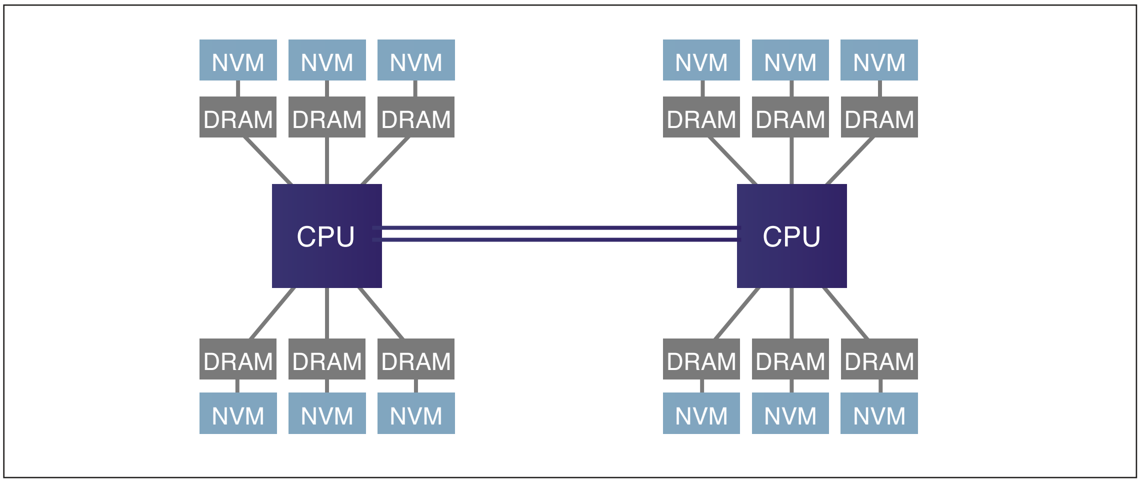

Figure 1: Main node DRAM / NVM (SCM) configuration.

Since we are also interested in maximum SCM bandwidth and capacity, we also need to populate all 12 memory channels in the node with SCM. Together, that makes 12 DRAM-DIMMs plus 12 NVDIMMs, pairing up in each channel. Of course, devices on the same channel share the available DDR4 channel's bandwidth.

The server hardware is implemented in a 1U form factor (Figure 2), which is a good compromise between configurability and density.

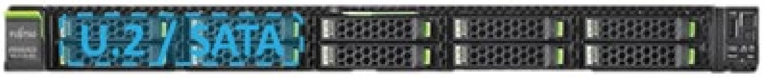

Figure 2: Unified NEXTGenIO server node (source: Fujitsu).

Figure 3 shows the NEXTGenIO server node implementation block diagram. While most details are not relevant here, some are important. In total, there are three PCIe x16 slots available for add-in cards in the system. We will develop support for processors with integrated network connection i.e. one channel coming off each CPU; this currently requires a PCIe slot to carry an "IFT Card", holding the modules to receive external cables (shown on the upper right in the picture). Consequently, that leaves two PCIe x16 slots for other I/O cards.

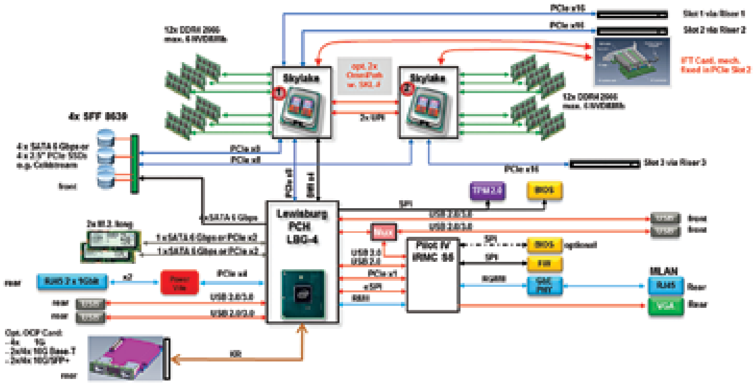

Figure 3: Server node block diagram.

3.1.2 Ethernet switch

For system management and general-purpose inter-node communication, two Ethernet networks are included: 1GE for the management LAN and slow data connections, and 10GE to connect to the data centre LAN as well as fast connections within the rack.

3.1.3 Network switch

The hardware architecture supports the integration of most common high-performance networks to provide the fast, high bandwidth, interconnect between the nodes in the system. This includes Infiniband based networks, Omni-Path Architecture and fast Ethernet and Omni-Path options, connected via PCI Express or (for Omni-Path) integrated into a multi-chip CPU package.

One feature that is necessary for a system to support all the project's requirements is remote data access over the network. RDMA offers the potential for high performance access to remote SCM, which is important to enable some of the use cases we are considering.

3.1.4 Login nodes

As with any HPC cluster, we need to support login nodes to access the system, perform compilations and remote visualisation. Login nodes have different requirements to the compute nodes, and as such will be provisioned with high memory and processing capabilities, along with local storage to enable fast compilation of applications.

This local storage can be provided by SCM in the nodes or local SSDs, depending on the cost and availability of these components. SCM could provide the benefit of a large memory space for the login nodes (i.e. if it supported 2LM functionality).

3.1.5 Boot nodes

To enable booting of the compute nodes via the network, for easy administration of the operating system, and to install software on the compute nodes, we provide a number of boot nodes. These have 1GE network connections to the compute servers, connected through 10GE switches (to reduce the risk of becoming saturated during system initialisation).

Boot nodes typically have low requirements on compute performance, i.e. even a mainstream single CPU is expected to be sufficient. The storage capacity requirements are also moderate. However, to avoid bottlenecks on boot storms, SSDs or SHDDs (HDDs with an internal NAND-Flash buffer) seem to be appropriate.

3.1.6 Gateway nodes

If an Omni-Path network is used as the main interconnect within the system, and we need to connect the system to an external Lustre filesystem based on Infiniband, we will need some gateway servers that bridge between the two interconnects. Intel have techniques for building such gateways using standard Intel® Xeon® servers and Linux operating system.

3.2 Prototype considerations

The choice of network for the prototype is between Infiniband and Omni-Path based solutions. Whilst Infiniband offers hardware RDMA, Omni-Path provides the potential for a network integrated on to the processor (integrated OPA or iOPA). For the NEXTGenIO prototype we decided to evaluate the integrated network functionality of Omni-Path.

3.2.1 RDMA options for the prototype

There are a number of ways we can provide remote data access in the prototype system. If we utilised an Infiniband based network we could enable RDMA (supported by the network adapter hardware), providing very low latency (1 microsecond or less) and high bandwidth access, with small I/O operations. This hardware RDMA also does not use much of the compute resource in a node as the network adapter has its own processing hardware to support the RDMA options.

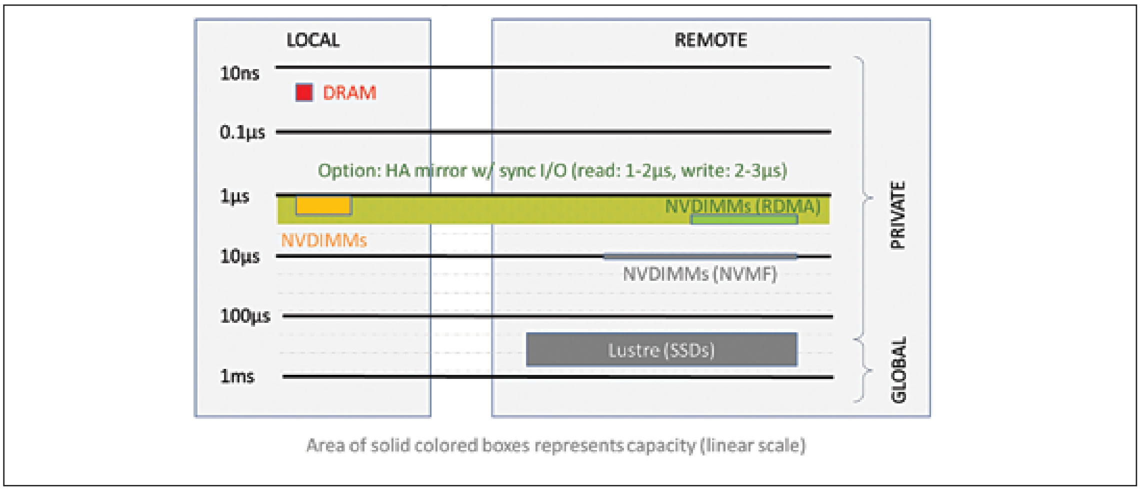
LOCAL
REMOTE
10ns
DRAM
0.1µs
Option: HA mirror w/ sync I/O (read: 1-2µs, write: 2-3µs)
1µs
NVDIMMs (RDMA)
NVDIMMs
10µs
NVDIMMs (NVMF)
100µs
Lustre (SSDs)
1ms
PRIVATE
GLOBAL
Area of solid colored boxes represents capacity (linear scale)

Figure 4: Local and RDMA SCM latency with hardware RDMA.

However, the current generation of Omni-Path network adapters does not support hardware accelerated RDMA (future generations of Omni-Path are scheduled to have such support). In this scenario, parts of the RDMA operations need to be executed by the host CPU. Assuming an optimised RDMA emulation target using polling, we expect a remote SCM block read/write latency of around 4 to 5 µs (Figure 5). Even though this is significantly higher than with hardware accelerated RDMA, it will still be in the range for effective use of synchronous I/O.

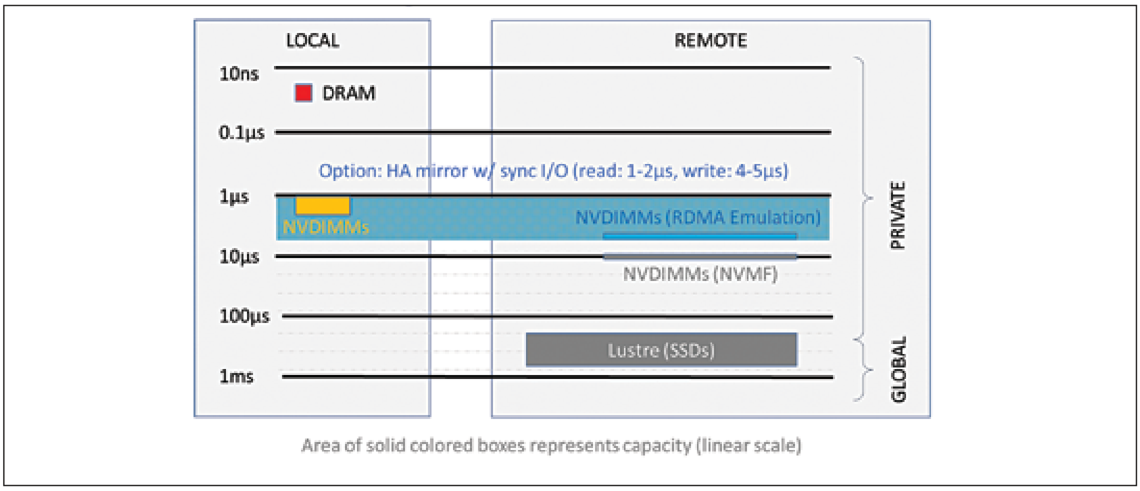
LOCAL
REMOTE
10ns
DRAM
0.1µs
Option: HA mirror w/ sync I/O (read: 1-2µs, write: 4-5µs)
1µs
NVDIMMs
NVDIMMs (RDMA Emulation)
10µs
NVDIMMs (NVMF)
100µs
Lustre (SSDs)
1ms
PRIVATE
GLOBAL
Area of solid colored boxes represents capacity (linear scale)

Figure 5: Local and remote SCMs with RDMA emulation.

3.2.2 Applicability of the architecture without access to SCM

Whilst this hardware architecture is designed around SCM, it is more generally applicable. For instance, local (in-node) SCM could be replaced by local non-volatile disk drives, using NVM-Express (NVMe) devices. These provide byte level access to persistent memory from applications, and are compatible with the pmem.io library functionality [3]. It would also be possible to replace in-node SCM with remote NVMe devices, using NVM over Fabrics (NVMF).

NVMe latency and bandwidth will not be of the same level of performance compared to the SCM we will use in our prototype architecture (i.e. 3D Xpoint™ memory), with best sustained NVMe latencies for current technology being around 100-300 µs, and bandwidth of around 2.5-3.5 GB/s (for a x4 PCIe connection), compared to 300-500ns and 8-10 GB/s for the SCM (per DIMM). The NVMF latency will be higher than with remote SCM accessed by RDMA. We expect around 10-20 µs (read/write) with 3D XPoint™-SSDs (Optane) and around 10 µs (read/write) with a SCM based NVMF target.

[5] https://github.com/daos-stack/daos

3.3 System configuration

The NEXTGenIO prototype rack provides approximately 100TB NVRAM capacity based on Intel's 3D XPoint™ technology [1]. It is accessible from any (and to any) CCN in less than 5 µs, with very high I/O bandwidth (almost 400 MIOPs).

The rack holds up to 32 Intel® Xeon® servers, also known as complete compute nodes (CCN), connected by Intel® Omni-Path Architecture fabric. Two login nodes support local compilation, visualisation, and also rack management. Two boot servers enable network boot, and two gateways to an external file system. The node configurations details and network connectivity are shown in Figure 7.

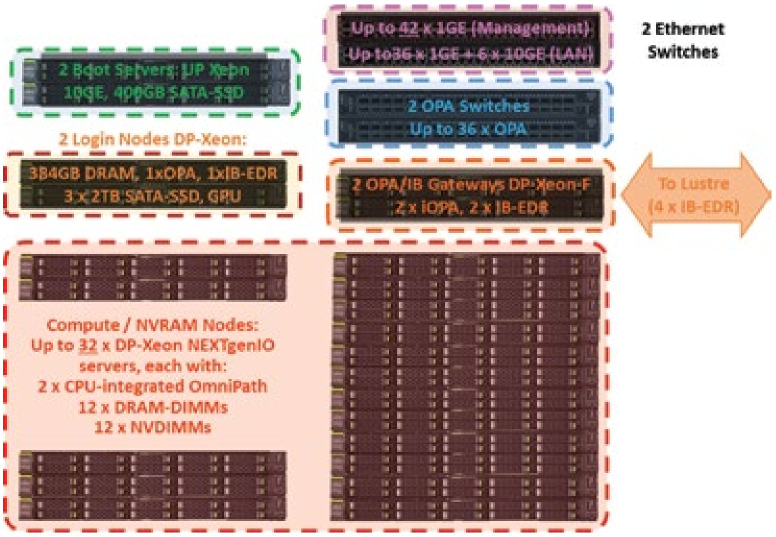

Figure 6: The NEXTGenIO prototype rack.

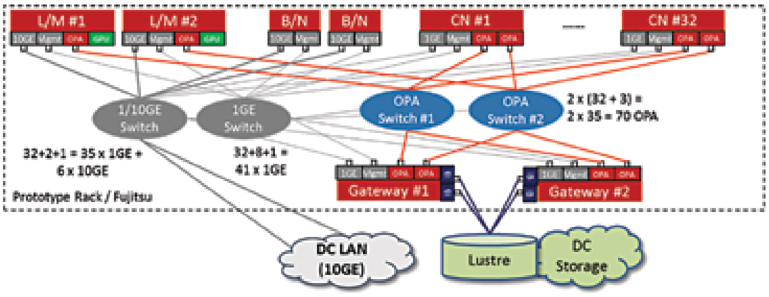

Figure 7: Prototype block diagram.

Assuming 2 to 4 TFLOPS per node, the NEXTGenIO rack with 32 nodes achieves 64-128 TFLOPS and approximately 384 MIOPS, i.e. around 1.5 TB/s local block storage. For the purpose of the prototype this scale is sufficient, but one key aspect of this work is to show how this could scale into the ExaFLOP range.

3.4 ExaFLOP vision

Since the architecture supports nodes using the two integrated network ports of the CPUs, we have two PCIe x16 slots available per node. Using two single-channel network cards, we can extend the single rack system to many racks, but still preserve the same hardware/software architecture. It would just add a second level of “slower” connectivity, i.e. between racks. “Slower” in this context means higher latency and lower bandwidth per node, but still keeping the concept of synchronous I/O, at least up to a range of 10 to 20 μs latency.

With a 5-hop network we can link 864 racks with 50 to 100 PFLOPs, 83 PB NVM, and 332 GIOPs:

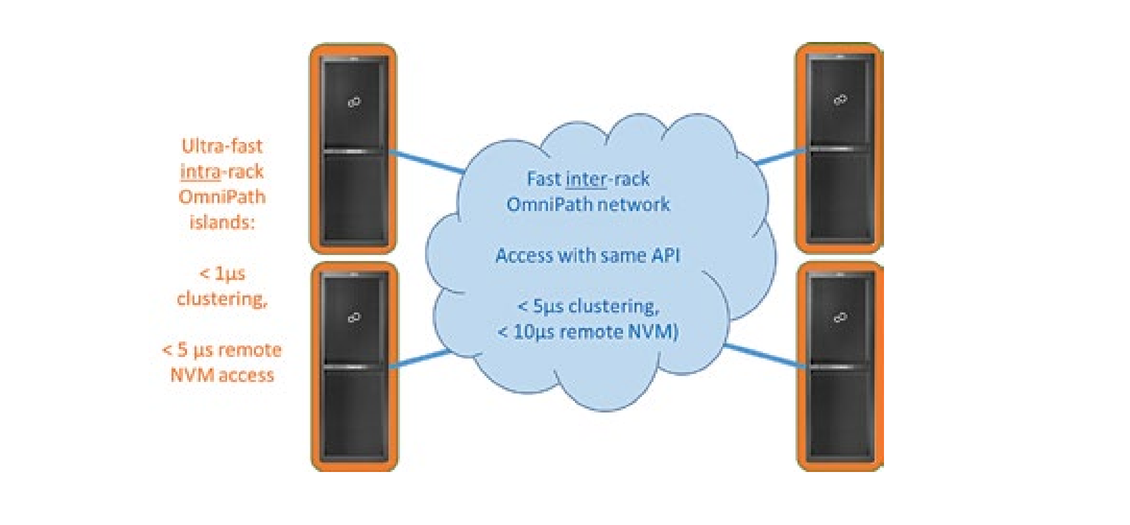

Figure 8: Scaling the NEXTGenIO architecture beyond one rack.

With cubical 3D torus configurations, we can project scaling to the ExaFLOP range – still using the same NEXTGenIO hardware/software architecture (Table 1).

Total # Nodes (Intel DP Skylake)	PFLOP (MIn Estimate)	PFLOP (Max Estimate)	Total SCM Capacity (PB)	Total SCM I/O B/W (TB/s)	Total Power (MW)
768	1.5	3	2	36	0.4
3.072	6	12	9	144	1.5
24,576	48	96	72	1,152	12
82,944	162	324	243	3,888	41
196,608	384	768	576	9,216	98
384,000	750	1,500	1,125	18,000	192

Table 1: NEXTGenIO scalability (3D torus projections).

At 750 to 1,500 PFLOPs, the total SCM capacity would be 1.1 Exabyte. The total I/O bandwidth would be 18 PB/s, which corresponds to the speed of approximately 6 million current day PCIe-based SSDs. The FLOP to I/O bandwidth ratio would be 1:50, i.e. at a similar level to the FLOP to memory bandwidth ratio that HPC systems aspire too.

The power needed for a 1 ExaFLOP system based on current Intel processors would be impractical – almost 200MW. However, based on future CPUs and interconnects, an ExaFLOP configuration within an affordable 20-30MW should become feasible early in the next decade.

3.4.1 Scalability to an ExaFLOP and beyond

Technically, the OPA link layer can address more than 10 million nodes [2]. Such cluster size would be far beyond any practical power limit (5 GW). Early next decade, however, a dual processor (DP) node may have 50 TFLOPs or more in the same power envelope. Then, 50,000 nodes would achieve 2.5 ExaFLOPs in a much more practical power budget of 25MW.

3.4.2 Scalability into 100 PFLOP/s

According to Intel, a 5-hop OPA cluster can link up to 27,648 nodes. With our NEXTGenIO rack and a single OPA inter-rack link per node, this would lead to the following configuration (see also Figure 9):

- 864 racks with 27,648 NEXTGenIO nodes
- 50 to 100 PFLOPs
- 83 PB of SCM with 332 GIOPs @ 4kB = 1.36 PB/s local I/O
- 14 MW power (assuming 500W/node including infrastructure)

The inter-rack bandwidth would be dependent on the network topology. Not knowing the configuration details, we guess the bandwidth per node will be lower than the 2 x 100G available in our prototype rack. The latency, however, would be similar. Assuming 1 microsecond or less round-trip per hop, the MPI latency should stay below 5 µs, and remote SCM via RDMA emulation below 10 µs. This would enable synchronous remote I/O SCM storage access, but at somewhat lower efficiency than within the racks.

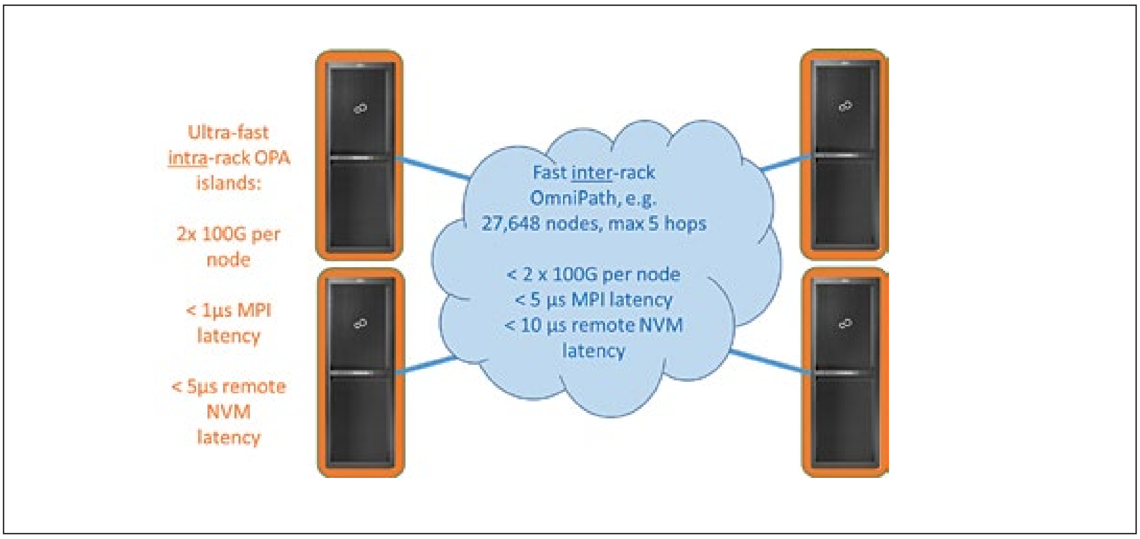
Ultra-fast intra-rack OPA islands:
2x 100G per node
< 1µs MPI latency
< 5µs remote NVM latency
Fast inter-rack OmniPath, e.g. 27,648 nodes, max 5 hops
< 2 x 100G per node
< 5 µs MPI latency
< 10 µs remote NVM latency

Figure 9: Intra-rack vs. inter-rack connectivity.

Since the bandwidth per node may be limited by this inter-rack bandwidth restrictions, the software should be aware of this extra network level. The scheduler and applications should therefore ensure that inter-rack accesses are less frequent than intra-rack accesses. This is very similar to today's NUMA memory access optimisation within compute nodes.

4 References

[1] "3D XPoint™ Technology Revolutionizes Storage Memory"
http://www.intel.com/content/www/us/en/architecture-and-technology/3D-XPoint™-technology-animation.html

[2] Intel Omni-Path Whitepaper from Hot Interconnects 2015: https://plan.seek.intel.com/LandingPage-IntelFabricsWebinarSeries-Omni-PathWhitePaper-3850

[3] pmem.io: http://pmem.io/