NextGenIO Workshop @ ECMWF System Architecture

Reading

Presenter: Bernhard Homölle

September, 2019

The Prototype
Overview

2 x 19” 42U Racks

34 x Complete Compute Nodes

1632 Cores /3264 Threads

192GB/6.37 TB DRAM

3TB/ 102TB DCPMM

2 x Login Nodes

2 x Boot Nodes

2 x Storage Nodes

1 x external SAS Storage Bay
As LUSTRE Storage

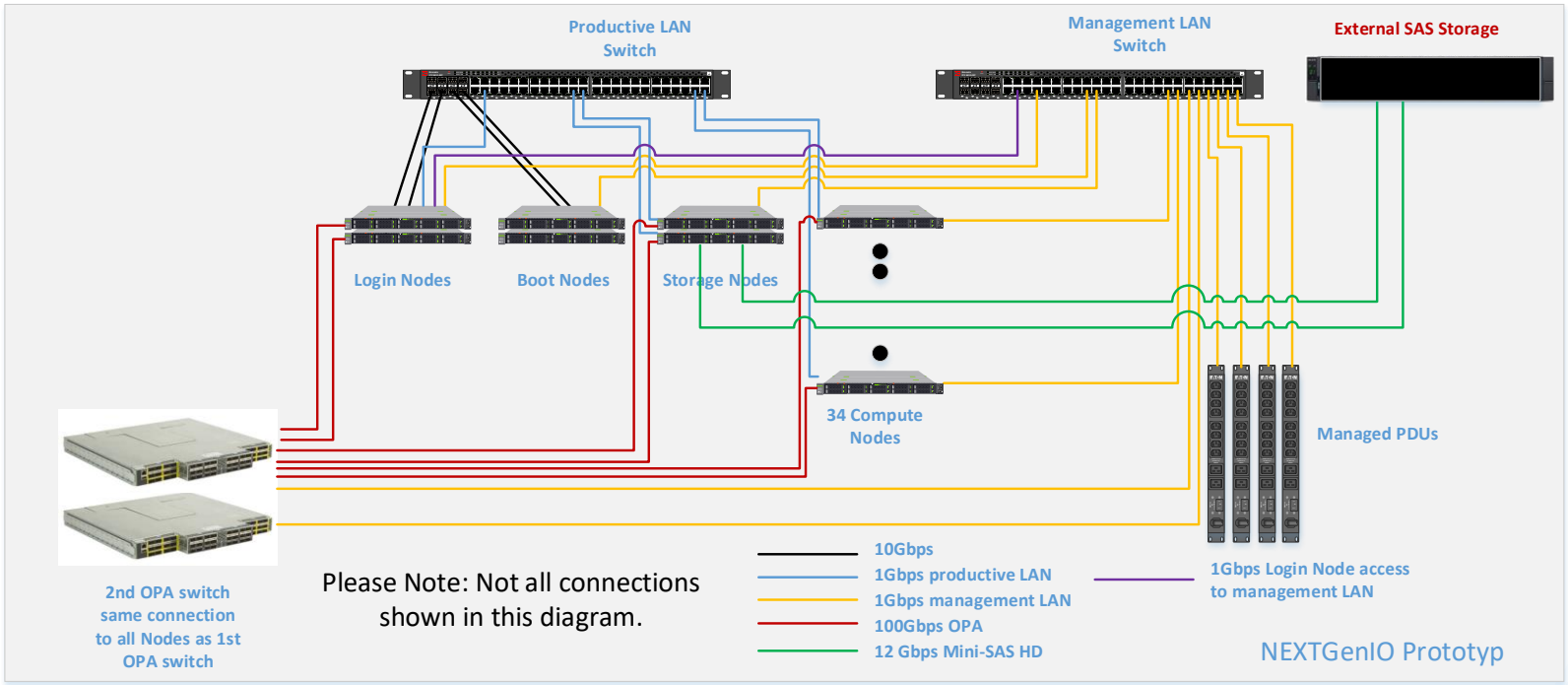
2 x 100Gbps Omni-Path Switch

1 x Management Network

1 x Productive LAN

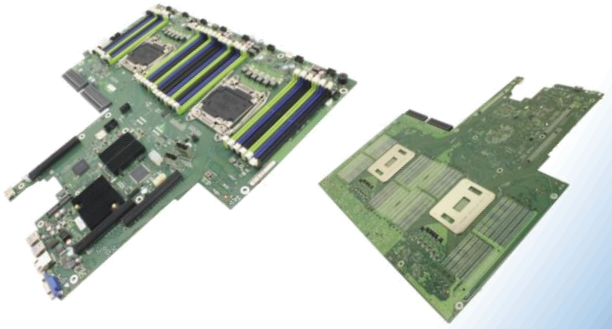
4 x managed PDUs

The Prototype
Infrastructure Topology view



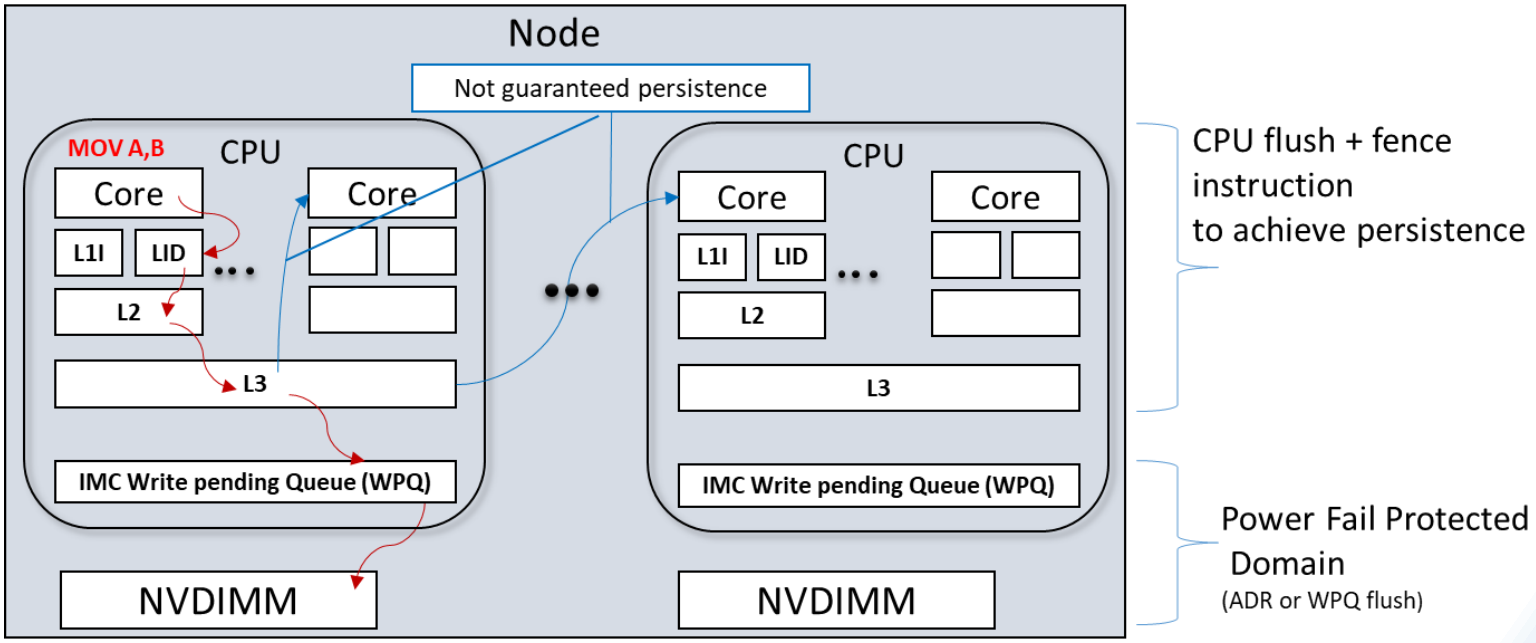
NEXTGenIO Motherboard
Developed @Fujitsu Augsburg

- Common Server Node
  - 1U 2 Socket system, Supported CPUs up to 205W
  - 24 DIMM slots / 12 DRAM + 12 NVDIMM
  - Up to 6TB Main Memory
  - 3x PCIe x16 Gen3 slots
  - 2x 1GbE onboard, 1 x 1GbE Mgnt LAN
- Motherboard
  - 12 Layers, 2 x LGA CPU sockets with 3647 pins
  - Optional FPGA on Die support
  - Optional Omni-Path On-Die Support
- UEFI BIOS
  - The BIOS project consists of about 15000 files
  - About 3000 files developed/overridden by Fujitsu
- iRMC S5
  - Board Management Controller
  - Firmware developed and modified for NVRAM



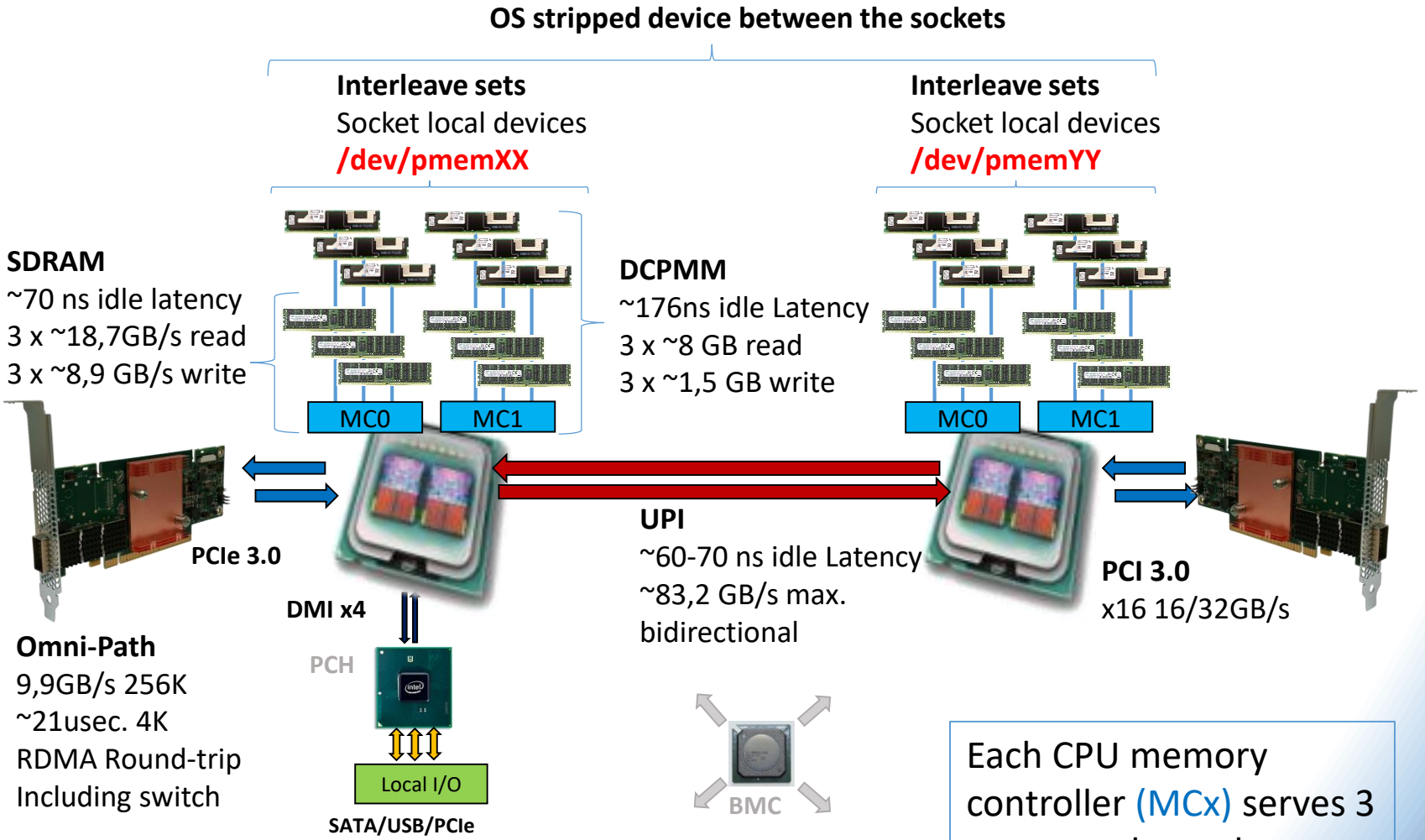
Platform support for ADR
Asynchronous DRAM Self Refresh

- Use energy stored in the PSU to protect CPU memory controller write pending queue (WPQ).
  - PSU send early AC-Fail, remaining time with stable voltage is used to drain the data from the WPQ to the NVDIMM.
  - NVDIMM will afterwards switch to Self Refresh to avoid false writes due to floating signals before total power down.



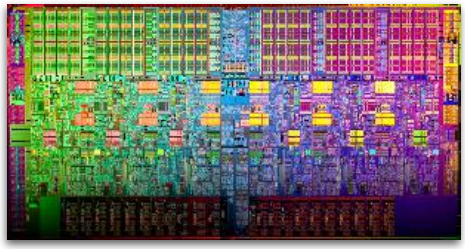
Power Fail Domain is just the WPQ and the DCPMM. Application has to flush the data out of the cache to make them persistent. Performance penalty. Fence operation is required to ensure ordering.

Basic Throughput & Latency



Each CPU memory controller (MCx) serves 3 memory channels.

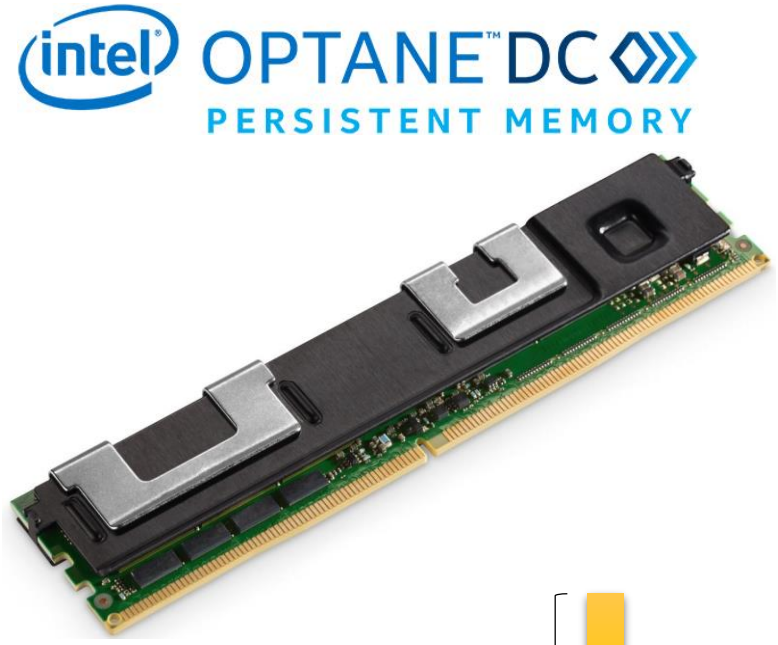
NEXTGenIO Prototype System
Intel Cascade-Lake Platinum 8260M CPU

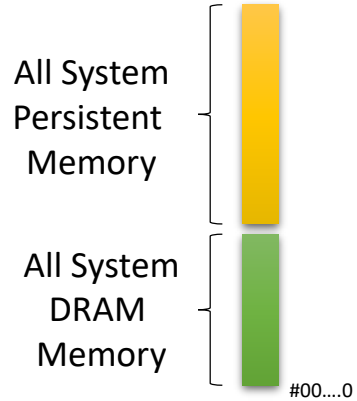


- 14 nm, 24 Cores @2.4GHz 165W, Turbo up to 3.9GHz, 2 x AVX 512
- 2TB memory max. with 2 integrated memory controller
- 6 Memory channel per socket with up to ~141GB/s burst rate total
- Support for up to 6 DCPMM per socket ⇔ 1.5TB NVRAM per socket

- L0 µOP cache: 1,536 µOPs/core, 8-way set associative
- L1I Cache: 32 KiB/core, 8-way set associative
- L1D Cache: 32 KiB/core, 8-way set associative Write-back policy
- L2 Cache: 1 MiB/core, 16-way set associative Write-back policy
- L3 Cache: 35.75MB 1.375 MiB/core, 11-way set associative, shared across all cores WB policy
- ITLB per Core
  - 4 KiB page translations: 128 entries; 8-way set associative
  - 2 MiB / 4 MiB page translations: 8 entries per thread; fully associative
- DTLB per core
  - 4 KiB page translations: 64 entries; 4-way set associative
  - 2 MiB / 4 MiB page translations: 32 entries; 4-way set associative
  - 1G page translations: 4 entries; 4-way set associative
- STLB per Core
  - 4 KiB + 2 MiB page translations: 1536 entries; 12-way set associative.
  - 1 GiB page translations: 16 entries; 4-way set associative

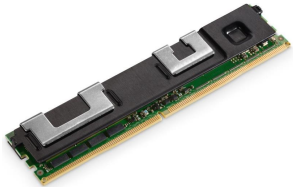
Intel® Optane™ DC (DCPMM)



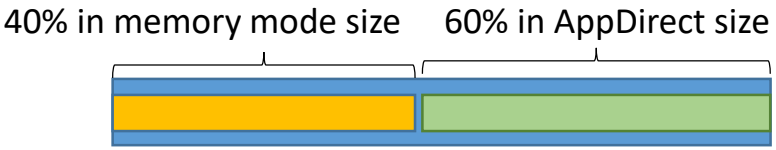


- Byte addressable connected to the memory bus like normal DRAM
- Cache Coherent, Load /Store Access
- DRAM-like Performance with low latency ~100 - ~300 nsec (normal operation).
- 128/256/512GB per module, 2 socket system 6TB
- Ability to do DMA & RDMA
- High Endurance with 5 years lifetime assuming maximum write bandwidth
  - ~500 Petabyte written per module (flash devices ~500-700 Terabyte)
- Turn time consuming **classic storage I/O into fast persistent memory** operation
  - No Paging for storage I/O
  - No Context Switching for storage I/O
  - No Interrupts for storage I/O
  - No Kernel Code Running for storage I/O

Setup DCPMM

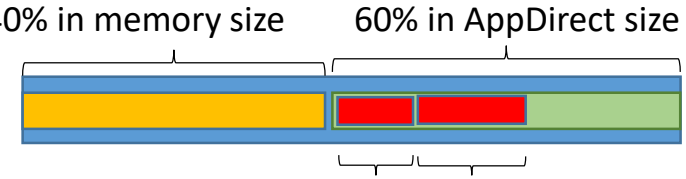


- Each DCPMM can be partitioned in two areas
  - **Memory mode space**
    can be used only in 2LM platform mode
  - **AppDirect space**, can be used in both 1LM and 2LM platform mode



Exampel:
#**ipmctl** create -goal MemoryMode=40

- Each AppDirect space can be divided into different namespaces
  - Namespaces are like LUNs
  - Can be formatted and mounted
  - Mounted with a dax (option –dax) capable file system direct access is possible. In that case the page cache is disabled!
  - Can be also mounted without dax option. In that case normal but fast block storage with page cache.



Exampel:
#**ndctl** create-namespace --mode fsdax --size 192G
Creates a block device under /dev/pmemxx
#**ndctl** create-namespace --mode devdax --size 192G
Creates a character device under /dev/daxxx

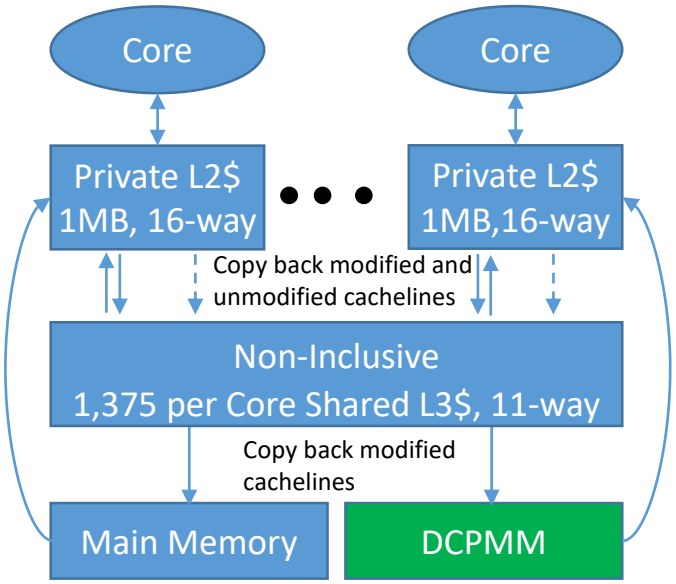
Example Mount:
#**mkfs.ext4** /dev/pmem3.1
#**mount -o dax** /dev/pmem3.1 /mnt/test/

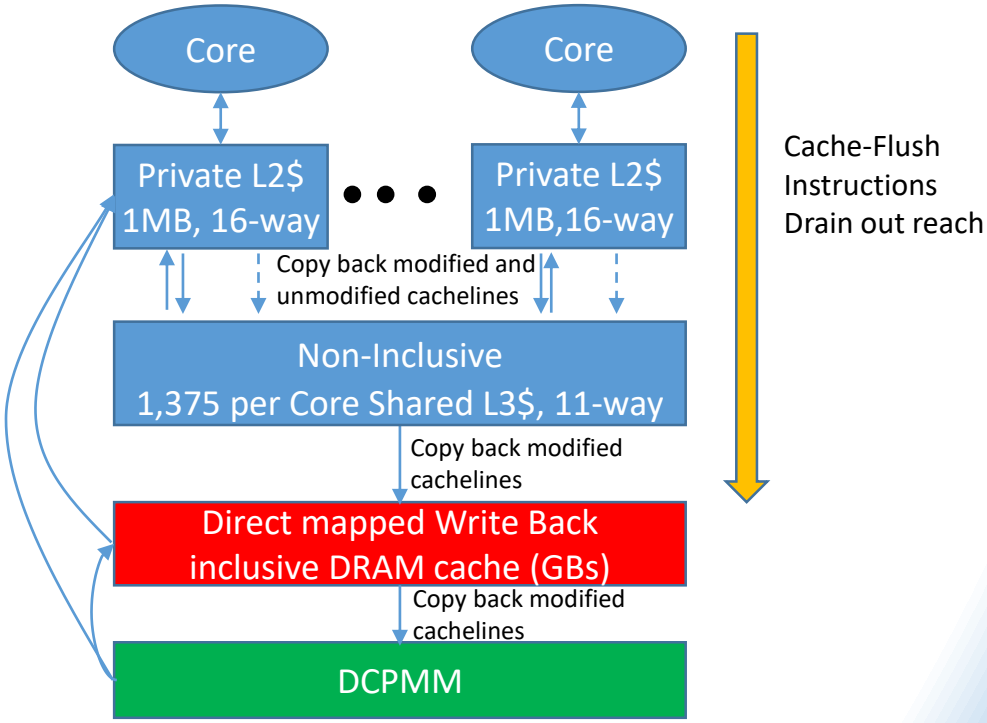
Platform Modes

- The NEXTGenIO Common Server Node support two different platform modes
  - The **1LM** mode:
    All DRAM is used for main memory
    Size of all DRAM DIMMs is the size of the OS main memory.
    NVRAM in AppDirect space can be access via memory mapped operations
  - The **2LM** mode:
    All DRAM is used as L4$ for the memory mode space.
    Size of all memory mode space in all DCPMM is the size of the OS main memory.
    AppDirect space (NVRAM) can be access via memory mapped operations

- To switch between the platform mode a reboot is required.
  - Platform mode is part of the system BIOS setting
    At least one DRAM per CPU memory controller is required to use DCPMM
  - In the NEXTGenIO Prototype 12 DRAM DIMMs and 12 DCPMM are built in.

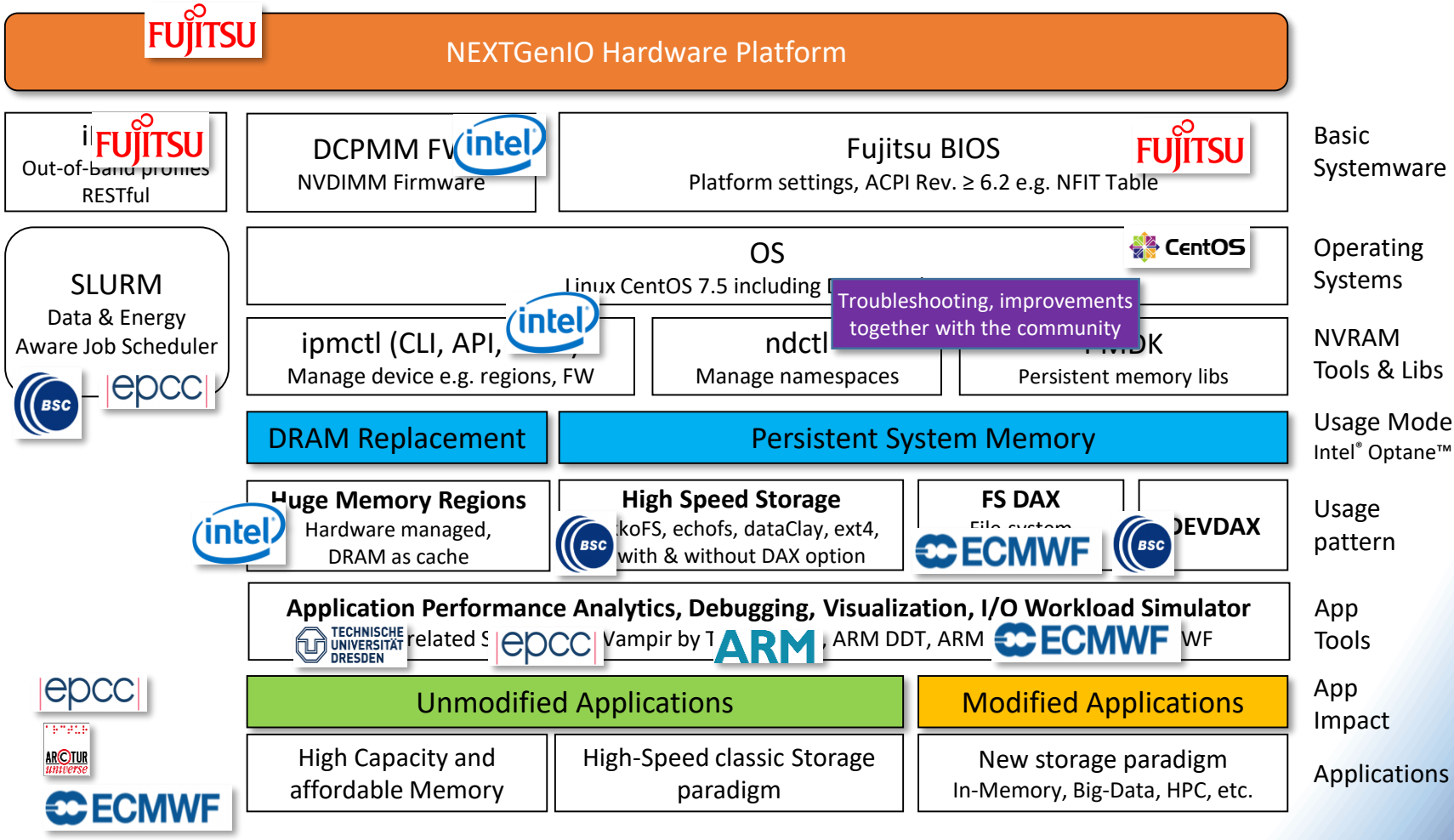
CPU Cache Hierarchy for 1LM and 2LM Mode





- Operation after Cache Miss
  - Data Flow

Solution Stack



Some Words on OS Pages and Caches

- The OS handles memory in blocks with continuous addresses of fixed length called pages.
  - Linus typical uses 4K page sizes
  - Modern CPUs support different page sizes such as 4K, 2MB or 1GB
  - Each page has a dirty bit to indicate that modified data are in that page.
  - Memory mapped file I/O **software** checks for dirty pages to sync/write them back.

- Caches are transparent to applications and OS
  - x86 uses 64 Byte cacheline (size of each entry)
  - The coherency protocol tags modified /dirty cachelines
  - Such dirty cachelines are automatically copy back by **hardware** to the memory if the space is needed (replacement).
  - Applications or OS can drain out cachelines if necessary.
    E.g. if persistence is required (cflush operations).

What Matters Pages or Caches?

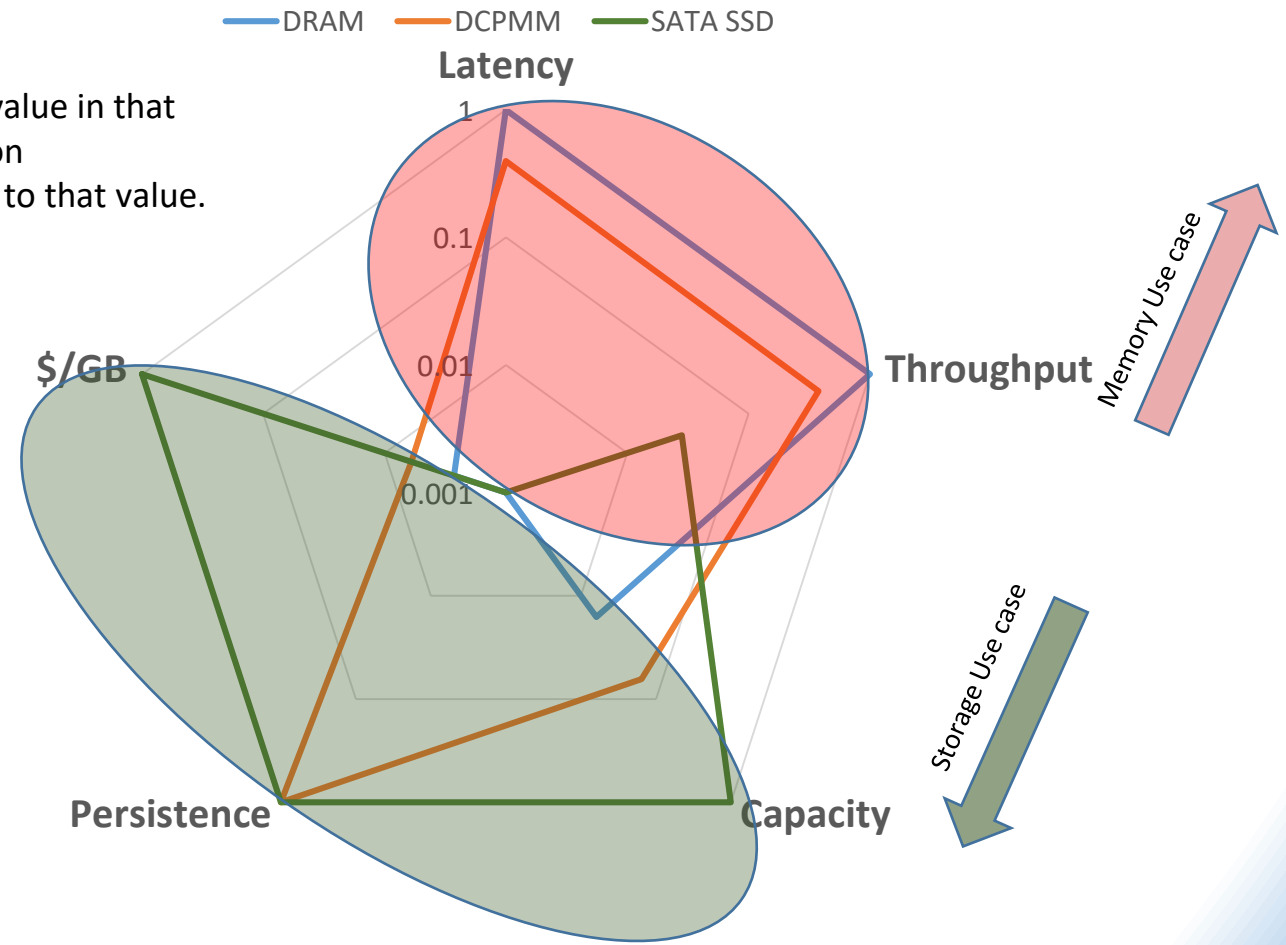
- Classic Memory Mapped File I/O persistence operations (msync) are directly bound to dirty pages.
- With DCPMM persistence operations are bound to dirty cachelines.

A cacheline is just 64Byte while pages are at least 4096Bytes!

- For DCPMM and Memory mapped files the page size is still important.
  - The good of big page sizes; Less table walks for huge files
  - The bad; Bigger pages can waste memory or storage (DCPMM) and can lead to fragmentation.

Comparison Memory & Storage

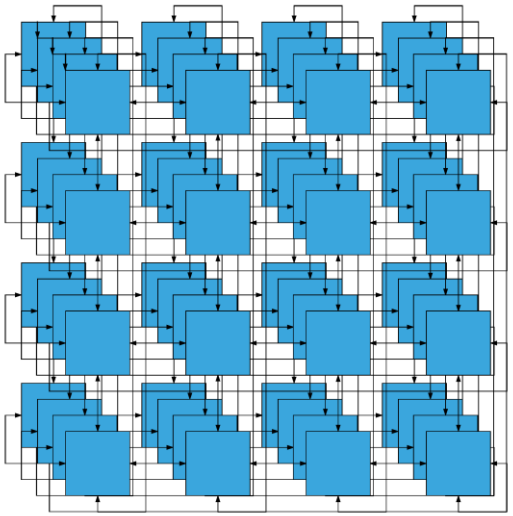
Normalized view
1 is for the best value in that particular function
Others a relative to that value.
128GB LRDIMM
8TB SSD
512GB DCPMM



Backup

Scaling the NEXTGenIO Architecture beyond one Rack

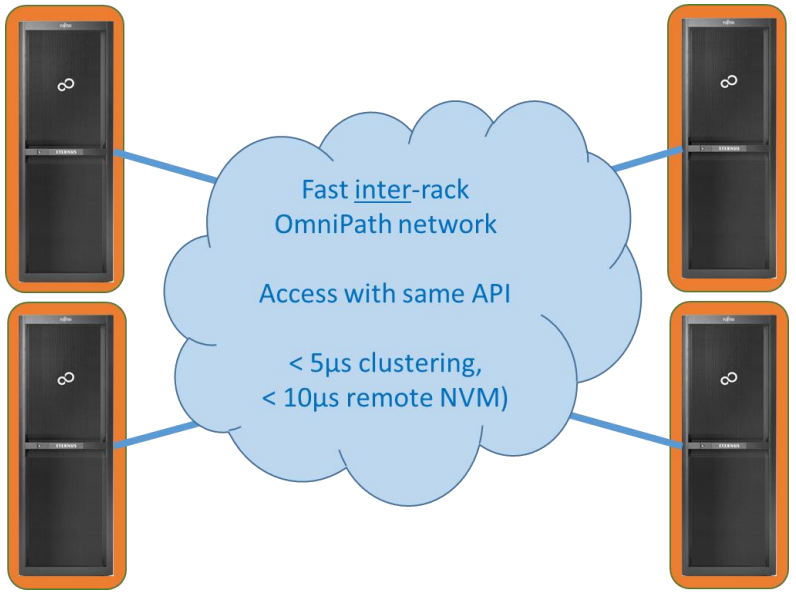
- With cubical 3D-torus configurations, we can project scaling to the ExaFLOP range



Ultra-fast intra-rack OmniPath islands:

< 1µs clustering,

< 5 µs remote NVM access



Total # Nodes (Intel DP)	PFLOP (Min Estimate)	PFLOP (Max Estimate)	Total NVDIMM Capacity (PB)	Total NVDIMM I/O B/W (TB/s)
768	1,5	2,2	2,3	36
3.072	6	9	9	144
24.576	48	70	72	1.152
82.944	162	235	243	3.888
196.608	384	557	576	9.216
384.000	750	1.088	1.125	18.000

*8260M 2 x AVX512, 12 x 256GB NVDIMM, 256B access pattern, no accelerator