HPC Solutions – Engineering Overview

HPC and AI Innovations Lab
www.hpcatdell.com

May 2021

DELL Technologies

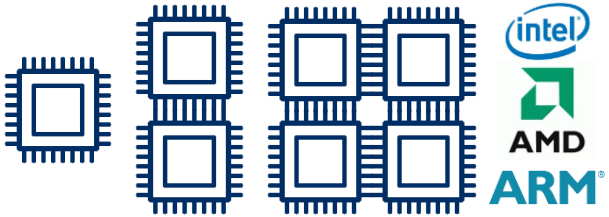
Dell Technologies HPC & AI Innovation Lab

Develop
Best Practices & Solutions

Industry-focused Research
with Customers and Partners

Contribute
to the Community

Customers Have LOTS of Questions



Which Processor? How Many?

ISILON
BeeGFS
NFS
PixStor

Which File System?

RoCE
Mellanox TECHNOLOGIES
intel
Intel® Omni-Path Architecture

Which Network?

NVIDIA

Do I Need Accelerators?

Should I Use Containers?

What Settings/Options/Flags?

DELL Technologies

World-class infrastructure in the Innovation Lab

13K ft.^{2} lab, 1,300+ servers, ~10PB storage dedicated to HPC and AI in collaboration with the community

Zenith

- TOP500-class system
- Was #383, #292, #265, #396 on Top 500
- 420 servers Xeon servers, HDR100 InfiniBand
- **~900 TF combined performance!**
- BeeGFS, Isilon H600 and Isilon F800 storage
- **Liquid** cooled and air cooled

Rattler

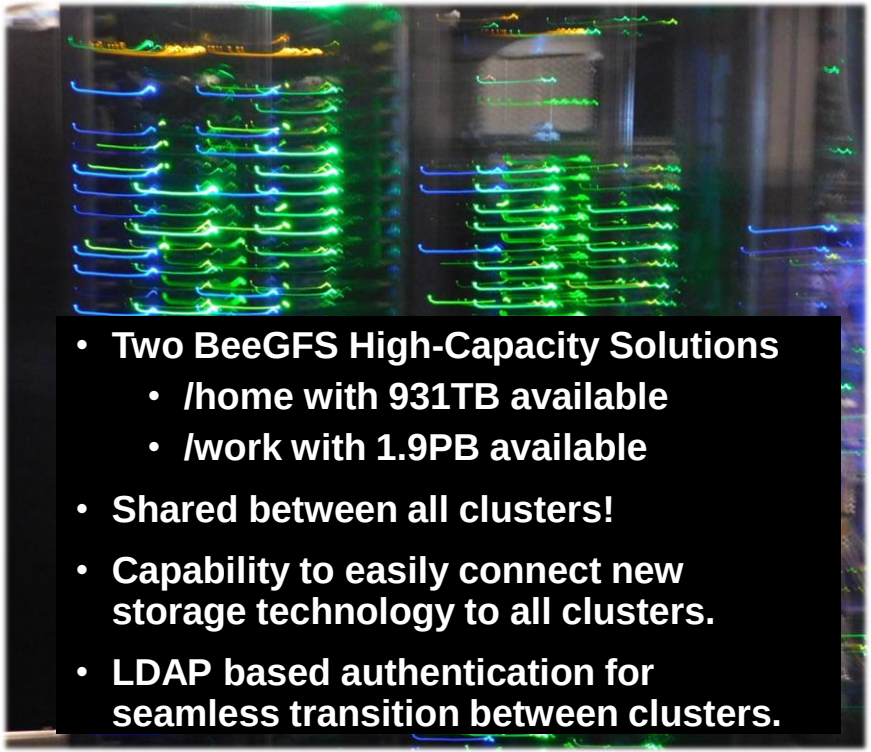
- Mellanox, NVIDIA and Bright Computing
- ~20 GPU cluster, HDR100 InfiniBand
- BeeGFS storage

Minerva

- 128 AMD EPYC Rome dual socket systems, HDR InfiniBand
- BeeGFS storage

Other systems

- Smaller test clusters, storage solutions, etc.



Intel Ice Lake Architecture

DELL Technologies

Ice Lake-SP IO and Memory Hierarchy

Integrating PCIe Gen4 controllers

- New IO Virtualization design, enables up to 3x BW scaling on large payloads (2x frequency, larger TLB, supports 2M/1G pages for in translation requests)
- New P2P credit fabric implementation to reach top P2P BW targets

3 independently clocked UPI links

4 Memory Controllers with enhanced per channel schedulers

- New memory controller design w/ optimizations

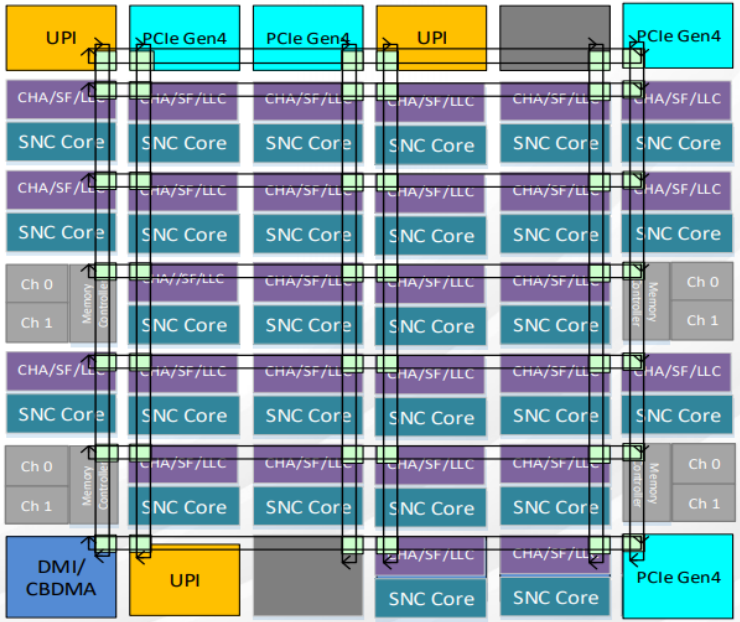
Intel® Total Memory Encryption (TME)

- DRAM encrypted using AES-XTS 128bit

Intel Optane Persistent Memory 200 Series (Barlow Pass)

- Higher speed and better power profile

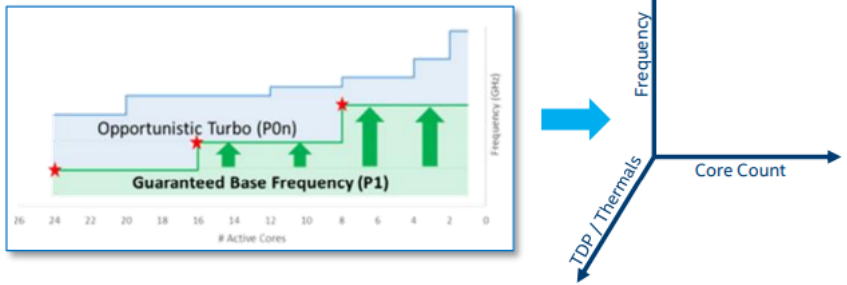
Ice Lake SP (28 core example)



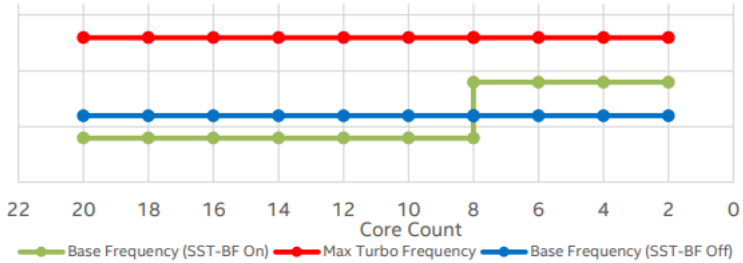
Intel® Speed Select Technology (Intel® SST) Features

Offers a suite of capabilities to allow users to re-configure the processor – dynamically, at runtime to match the usage / WL and maximize performance

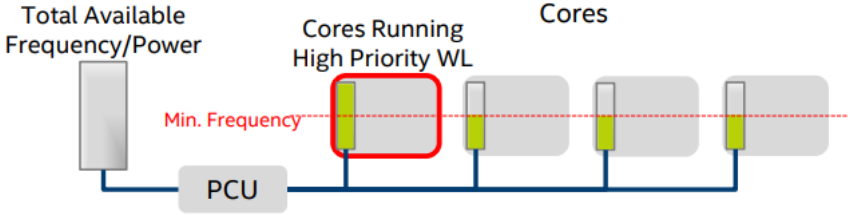
Intel Speed Select Technology–**Performance Profile** (Intel SST-PP)



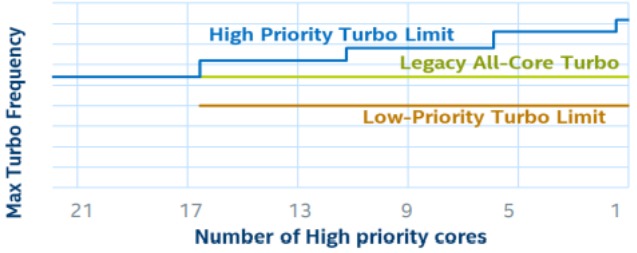
Intel Speed Select Technology–**Base Frequency** (Intel SST-BF)



Intel Speed Select Technology–**Core Power** Intel SST-CP



Intel Speed Select Technology–**Turbo Frequency** (Intel SST-TF)



The New Dell EMC PowerEdge Server Portfolio

Specialized– EDGE&TELCO

XR11

XR12

Specialized– GPU OPTIMIZED

R750xa

XE8545

RACK SERVERS

R550

R7515

R750xs

R750

R7525

R650

R6525

R650xs

R6515

R450

C-SERIES

C6525

C6520

MODULAR COMPUTE SLED

MX750c

ITALICS:
ALL NEW INTEL ICE LAKE 2-SOCKET SERVERS

Rectangle shape: Intel Ice Lake servers as part of HPC & AI Solutions - Bravo

YOUR INNOVATION ENGINE

Technology and solutions that help you innovate, adapt, and grow

Interconnects

Snoop Hold off – BIOS Option

Roll256Cycles

	WindowSize=64		WindowSize=512	
Message Size	Bandwidth (GB/s)	Messages/s	Bandwidth (GB/s)	Messages/s
1	0.0	14 M	0.1	105 M
2	0.1	40 M	0.3	130 M
4	0.5	118 M	0.8	188 M
8	0.9	116 M	1.3	162 M
16	1.7	107 M	3.1	191 M
32	3.9	121 M	4.8	149 M
64	0.7	11 M	2.6	41 M
128	1.1	9 M	1.4	11 M

Roll2KCycles

	WindowSize=64		WindowSize=512	
Message Size	Bandwidth (GB/s)	Messages/s	Bandwidth (GB/s)	Messages/s
1	0.2	156 M	0.2	204 M
2	0.3	160 M	0.4	195 M
4	0.6	160 M	0.8	191 M
8	1.3	158 M	1.5	188 M
16	3.0	189 M	3.1	191 M
32	4.6	144 M	4.8	149 M
64	4.5	71 M	4.7	73 M
128	7.3	57 M	8.2	64 M

- OSU Message rate test with all cores.
- Selects the number of cycles PCI I/O can withhold snoop requests, from the CPU.
- Additional SnoopHldOff options are being added to the next block BIOS releases.

Mellanox HDR - Gotchas

- SNAPI Only : `virt_enable` set to 2 in opensm.conf

- Achieve Full BW on SNAPI cards, C6520 server
  - `ADVANCED_PCI_SETTINGS` should be set to `TRUE`
  - To achieve full BW from local and remote socket `MAX_ACC_OUT_READ` should be set to `16` for SNAPI cards

- On going discussion with Mellanox on HDR200 BiBW

Numanode	Bandwidth at 4MB message size
0	39.2 GB/s
1	37.8GB/s
2	**49.2 GB/s**
3	41.3GB/s

```
[root@gpu105 ~]# mst status -v
MST modules:
------------
    MST PCI module is not loaded
    MST PCI configuration module loaded
PCI devices:
------------
DEVICE_TYPE          MST                        PCI      RDMA     NET              NUMA
ConnectX6(rev:0)     /dev/mst/mt4123_pciconf0   98:00.0  mlx5_0   net-ib0          2
```

Application performance and BIOS tuning for HPC

DELL Technologies

HPL (optimization and performance)

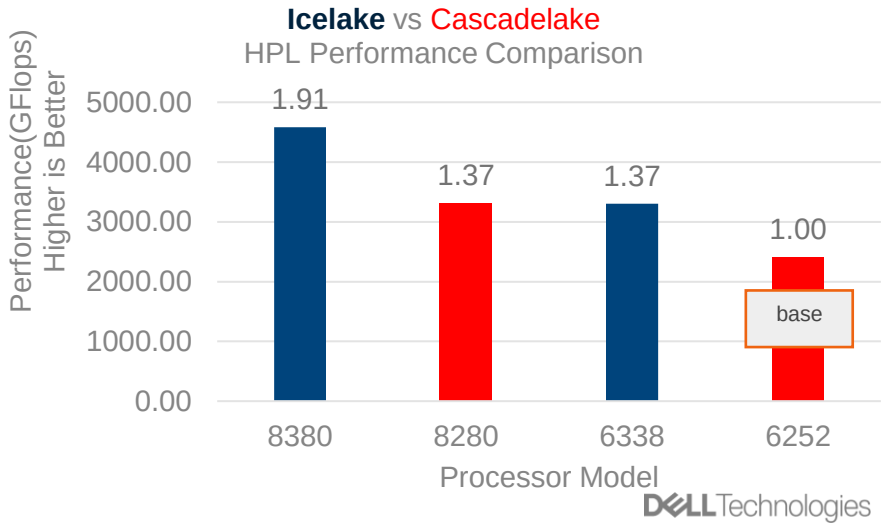
The best application performance can be achieved with the HPL setup bundled with Intel Parallel Studio.

In case of open-source version of HPL -

- Intel MKL is recommended
- Intel compiler **-qopt-zmm-usage=high -xICELAKE-SERVER** is the appropriate architecture flag, which enables AVX 512 SIMD instruction support

1 process per NUMA node is the recommended launch configuration

CPU	Cores	Frequency	Performance(GFlops)	Efficiency
8380	40	2.3 - 3.4 GHz	4586.80	0.78
6338	32	2.0 - 3.2 GHz	3304.64	0.81
8280	28	2.7 – 4.0 GHz	3308.06	0.68
6252	24	2.1 - 3.7 GHz	2407.13	0.68



STREAM Dual Socket(optimization and performance)

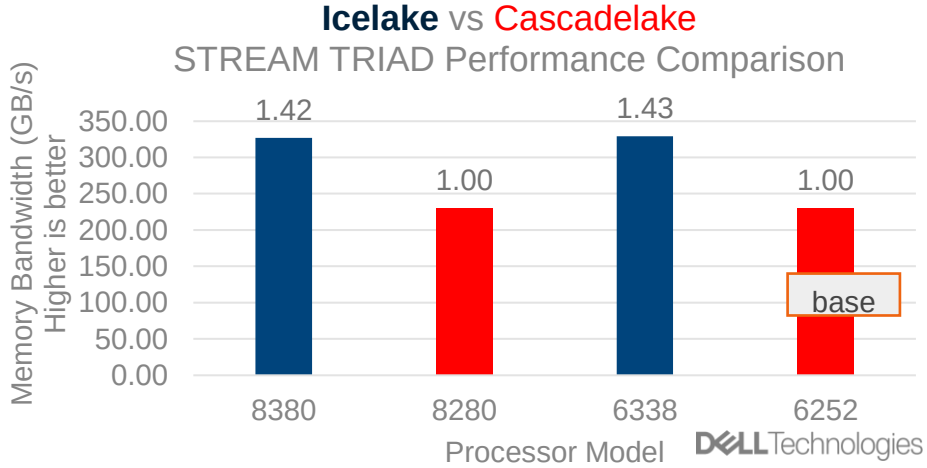
Intel compilers are recommended to get expected performance.

Streaming/non-temporal store support is required for optimal performance numbers

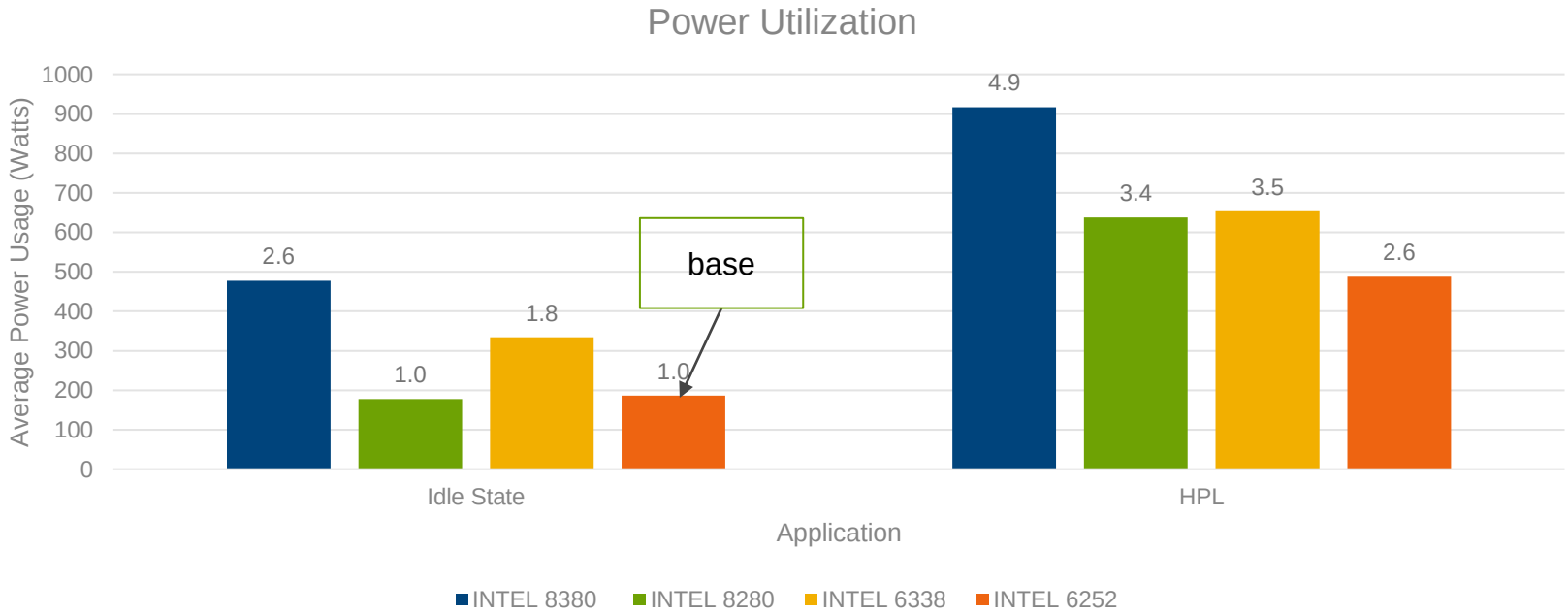
Recommended compiler flags (intel compiler) -
-xICELAKE-SERVER -O3 -ffreestanding -qopenmp -qopenmp-link=static -mcmodel=medium -shared-intel -restrict -qopt-streaming-stores **always** -DSTREAM_ARRAY_SIZE=160000000 -DNTIMES=100 -DOFFSET=0 -DVERBOSE -qopt-zmm-usage=high

While running KMP_AFFINITY environment variable should be set to “granularity=fine,scatter” , following environment
The system file /sys/kernel/mm/transparent_hugepage/enabled should be set to never.
STREAM TRIAD results were generated by subscribing all available cores on system.

CPU	Cores	Frequency	Performance(GB/s)	Efficiency
8380	40	2.3 - 3.4 GHz	326.8	0.80
6338	32	2.0 - 3.2 GHz	328.9	0.80
8280	28	2.7 – 4.0 GHz	230.3	0.82
6252	24	2.1 - 3.7 GHz	230.5	0.82



Power Utilization – Icelake vs Cascadelake



CPU	Cores	Frequency	TDP
8380	40	2.3 - 3.4 GHz	270W
6338	32	2.0 - 3.2 GHz	205W
8280	28	2.7 - 4GHz	205W
6252	24	2.1 - 3.7GHz	150W

Open Issues w/ RHEL 8.3

Issue Sighting – ICX-SP C-States with RHEL 8.3

Issue Description

- The base RHEL 8.3 kernel 4.18.0-240.el8 does not include C-state definitions for Ice Lake in the intel_idle driver.
- This results in C-state behavior for Ice Lake that is not consistent with previous generation Intel processors and not consistent with patched kernels.

Resolution

- The intel_idle driver was patched in the 4.18.0-240.11.1.el8_3 update kernel to include Ice Lake C-state definitions.
- Recommend updating to the 4.18.0-240.11.1.el8_3 or later kernel.

Issue Identification

- **Base kernel uses ACPI c-states when C-states are enabled in BIOS.**
- **Patched kernel uses intel_idle defined C-states, which are always enabled by default.**

Base Kernel 4.18.0-240

```
$ cpupower idle-info
CPUidle driver: intel_idle

Number of idle states: 3
Available idle states: POLL C1_ACPI C2_ACPI
POLL:
Flags/Description: CPUIDLE CORE POLL IDLE
Latency: 0
C1_ACPI:
Flags/Description: ACPI FFH INTEL MWAIT 0x0
Latency: 1
C2_ACPI:
Flags/Description: ACPI FFH INTEL MWAIT 0x20
Latency: 41
```

Patched Kernel 4.18.0-240.22.1

```
$ cpupower idle-info
CPUidle driver: intel_idle

Number of idle states: 4
Available idle states: POLL C1 C1E C6
POLL:
Flags/Description: CPUIDLE CORE POLL IDLE
Latency: 0
C1:
Flags/Description: MWAIT 0x00
Latency: 1
C1E:
Flags/Description: MWAIT 0x01
Latency: 4
C6:
Flags/Description: MWAIT 0x20
Latency: 128
```

C-State Influence on Turbo Frequency Behavior

Processor cannot reach maximum turbo frequency without C-states

C-States Disabled

2.8 GHz, 32 threads
2.8 GHz, 30 threads
2.8 GHz, 28 threads
2.8 GHz, 26 threads
2.8 GHz, 24 threads
2.8 GHz, 22 threads
2.8 GHz, 20 threads
2.8 GHz, 18 threads
2.8 GHz, 16 threads
2.8 GHz, 14 threads
2.8 GHz, 12 threads
2.8 GHz, 10 threads
2.8 GHz, 8 threads
2.8 GHz, 6 threads
2.8 GHz, 4 threads
2.8 GHz, 2 threads
2.8 GHz, 1 thread

C-States Enabled

2.8 GHz, 32 threads
2.8 GHz, 30 threads
2.8 GHz, 28 threads
3.0 GHz, 26 threads
3.1 GHz, 24 threads
3.1 GHz, 22 threads
3.2 GHz, 20 threads
3.3 GHz, 18 threads
3.4 GHz, 16 threads
3.4 GHz, 14 threads
3.4 GHz, 12 threads
3.4 GHz, 10 threads
3.4 GHz, 8 threads
3.4 GHz, 6 threads
3.4 GHz, 4 threads
3.4 GHz, 2 threads
3.4 GHz, 1 thread

- Active cores frequency behavior for Intel Xeon Platinum 8352Y

NVIDIA GPUs

DELL Technologies

RTS: May 2021

R750xa

CPU	2U2S, Ice Lake (PCIe Gen4)
Memory	32x DDR4 3200 MT/s
GPU/ FPGA	Offering the latest GPUs by NVIDIA: A100, A40 with NVLINK Bridges, M10 and T4; and AMD: MI100
Storage	Up to 8 SAS/SATA SSD or NVMe drives Optional BOSS
I/O	Up to 8 x PCIe Gen4 Slots (6 x16, 2 x8)
Network	2x1GbE LOM; 1x8 Gen4 OCP3.0
Cooling	High Performance Fans Optional Liquid cooling support for CPUs
PSU	1+1 1400W, 2400W (Platinum)

R750xa Value Proposition

PowerEdge go-to platform for GPU-optimized workloads

The R750xa offers:

GPU Optimization

- Massive compute power for the most complex accelerator workloads
- **Intel Ice Lake CPU and PCIe 4.0** to unleash the full capabilities of GPU-base compute

Workload Flexibility

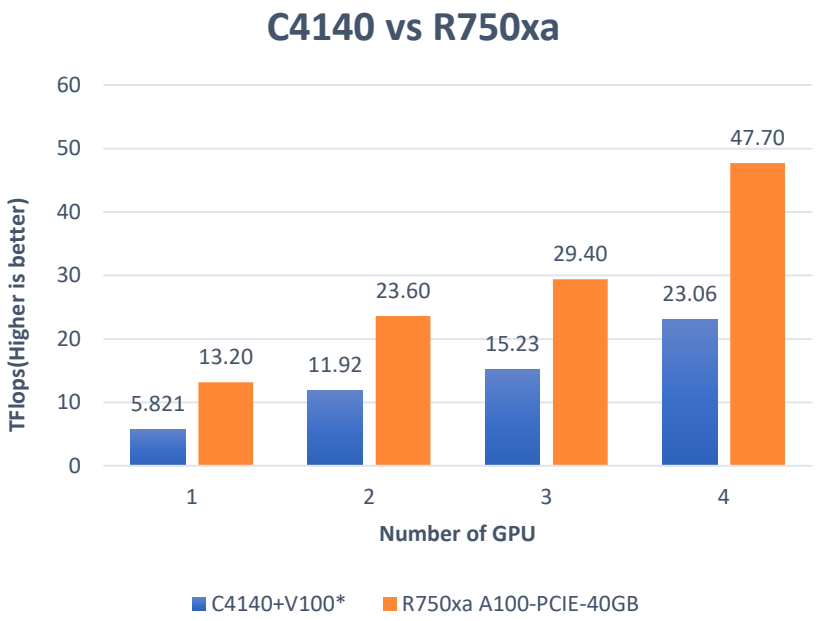
- Full-featured **support for the complete stack of GPUs in the PowerEdge portfolio**
- Max performance for the entire spectrum of **HPC, AI-ML/DL training and inferencing, DB Analytics and VDI workloads**

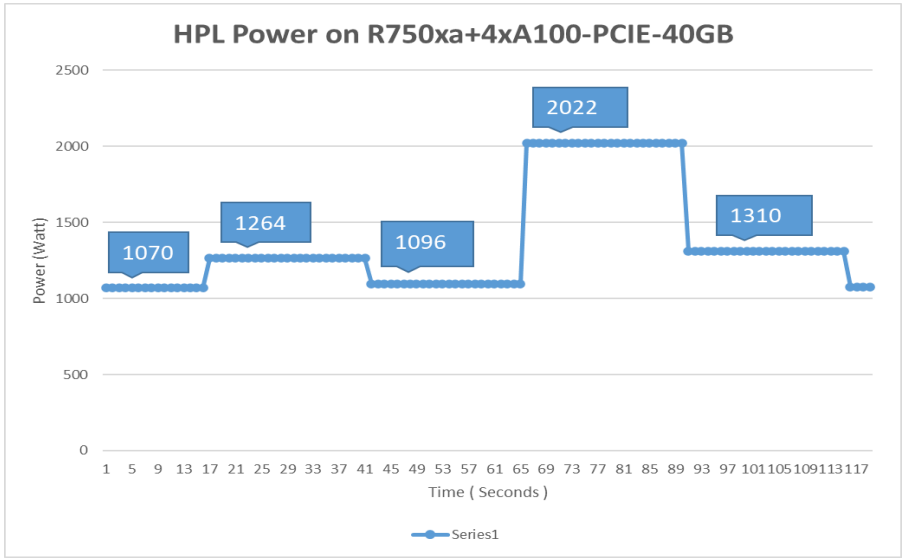
Scalable Density

- **Air cooled 2U** with ambient temperature of up to 35C
- GPU density of 2GPU/U with additional **support for the newly introduced NVLINK Bridges**
- Optional liquid cooling for CPUs to capture up to 20% of heat dissipation

DELL Technologies

HPL

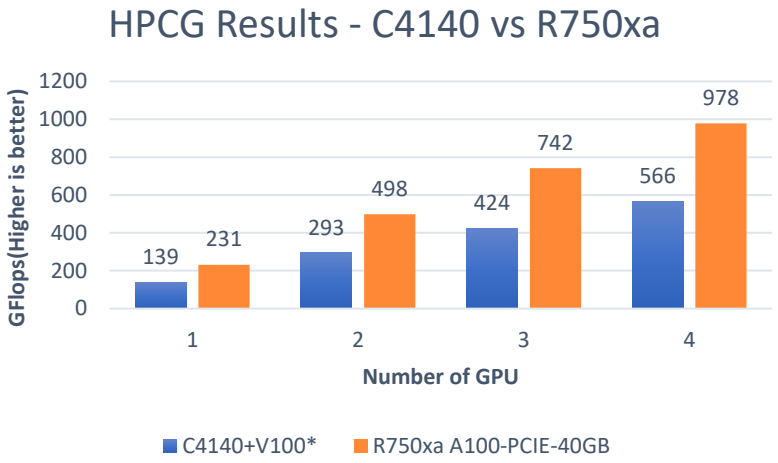


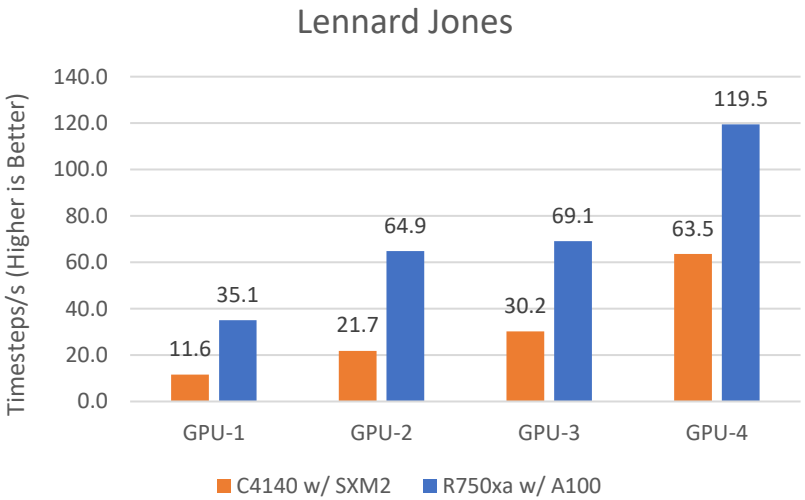


- ~2x performance with A100.
- Higher double precision value with A100
- Large problem size

Questions ?


DELL Technologies





- Memory bandwidth dependent
- 900GB vs 1555GB
- ~1.7x improvement at 4-GPUs

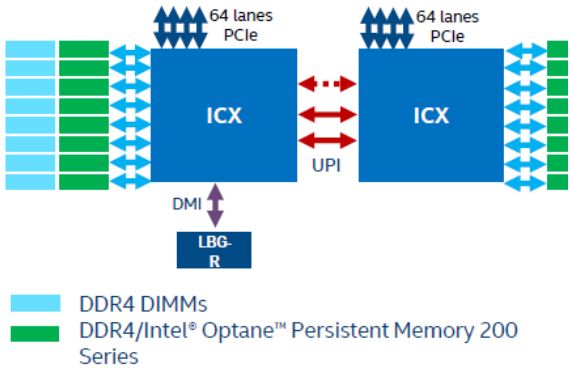
- LAMMPS is double precision application.
- Up to 2x performance improvement with A100 GPUs

Whitley/Ice Lake 2S Overview

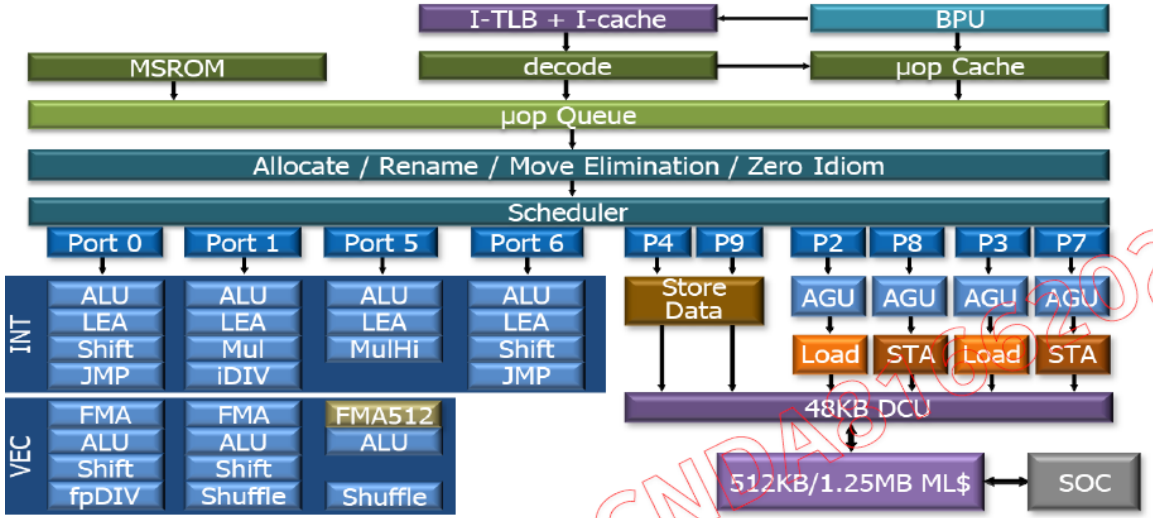
All SKUs, frequencies, and performance estimates are PRELIMINARY and can change without notice.

CPU	**Ice Lake (up to 270W[1])** **52b/57b Physical Address/Virtual Address**
New Capabilities	Intel® DL boost: VNNI for inference. **No support for BFLOAT16** **Crypto Enhancements: 2xAES, SHA Extensions, VPMADD52** **Database compression: VBMI** **Security: Intel® TME/TME-MT, Intel® SGX, PFR**
Socket	**Socket P+ 4189 pin**
Scalability	1S, 2S
Memory	8 channels DDR4 per CPU @3200[3] 2DPC, 16 DIMMs per socket **New Intel® Optane™ Persistent Memory 200 Series module[4]**
Intel® Ultra Path Interconnect (Intel® UPI)	Up to 3 links per CPU x20, speed: 10.4 and **11.2 GT/s**
PCIe	**PCIe 4.0: Up to 64 lanes per CPU[2]** (bifurcation support: x16, x8, x4) **up to 48 lanes on North, 16 lanes on South of socket,** NTB
PCH – Intel® C620A Series Chipset **(LBG-R)**	IE, Intel QAT, eSPI, **No Integrated 4x10GbE/1GbE ports**, Legacy 1GbE for manageability support, Up to 14 SATA 3, Up to 14 USB 2.0, Up to 10 USB 3.0, Up to 20 ports PCIe* 3.0 (8 GT/s) **Enhanced security through hardware via new stepping**

Whitley 2S Configuration



Sunny Cove Core Microarchitecture



	Cascade Lake (per core)	Ice Lake (per core)
Out-of-order Window	224	384
In-flight Loads + Stores	72 + 56	128 + 72
Scheduler Entries	97	160
Register Files – Integer + FP	180 + 168	280 +224
Allocation Queue	64/thread	70/thread; 140/1 thread
L1D Cache (KB)	32	48
L1D BW (B/Cyc) – Load + Store	128 + 64	128 + 64
L2 Unified TLB	1.5K	2K
Mid-level Cache (MB)	1	1.25

- Improved Front-end: higher capacity and improved branch predictor
- Wider and deeper machine: wider allocation and execution resources + larger structures
- Enhancements in TLBs, single thread execution, prefetching
- Server enhancements – larger Mid-level Cache (L2) + second FMA

Ice Lake & Cooper Lake Product Numbering Convention for Intel® Xeon® Scalable Processors

SKU Level
- 9, 8 = Platinum
- 6, 5 = Gold
- 4 = Silver
- 3 = Bronze

Processor Generation
- 1 = 1st Gen (Skylake)
- 2 = 2nd Gen (Cascade Lake & Cascade Lake-R)
- **3 = 3rd Gen (Ice Lake (2S), Cooper Lake (4+S))**

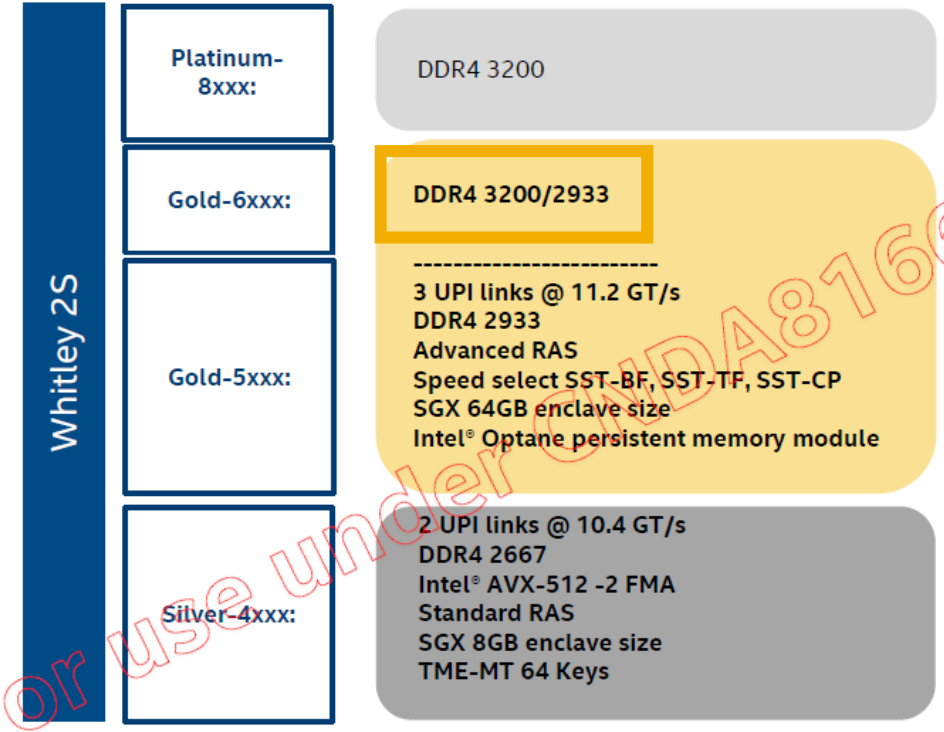
Processor SKU
- (ex. 20, 34…)

Integrations and Optimizations
(when applicable)
- H = Cooper Lake 4/8-socket with 6x UPI
- N = NFV Optimized
- T = High Tcase
- U = Single Socket
- V = SaaS optimized SKU for orchestration efficiency targeting high density, lower power VM environment (70% CPU utilization)
- P = IaaS optimized SKU for orchestration efficiency targeting higher frequency for VM Markets (70% CPU utilization)
- Y = Intel® Speed Select Technologies (PP, BF, TF, CP)
- S = max SGX enclave size SKUs (512GB)
- Q = Liquid cooling (Temperature Inlet to cold

Memory Capacity per socket
- No Suffix = 1TB/Socket memory tier
- L = 4.5TB/Socket memory tier

Note: **All information provided here is subject to change without notice.** Intel may make changes to specifications and product descriptions at any time, without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps. For latest information please refer to the Snapshot

Preliminary Guidance – Example SKUs Included as Reference



Bold indicates introduced in shelf

All SKUs, frequencies, and performance estimates are PRELIMINARY and can change without notice.

Note that not all Gold 63xx SKUs support 3200 MTps DIMM speeds.