HPC Technology Preview

Deepthi Cherlopalle
HPC and AI Innovations Lab
www.hpcatdell.com

May 2021

Dell Technologies HPC & AI Innovation Lab

Develop
Best Practices & Solutions

Industry-focused Research
with Customers and Partners

Contribute
to the Community

World-class infrastructure in the Innovation Lab

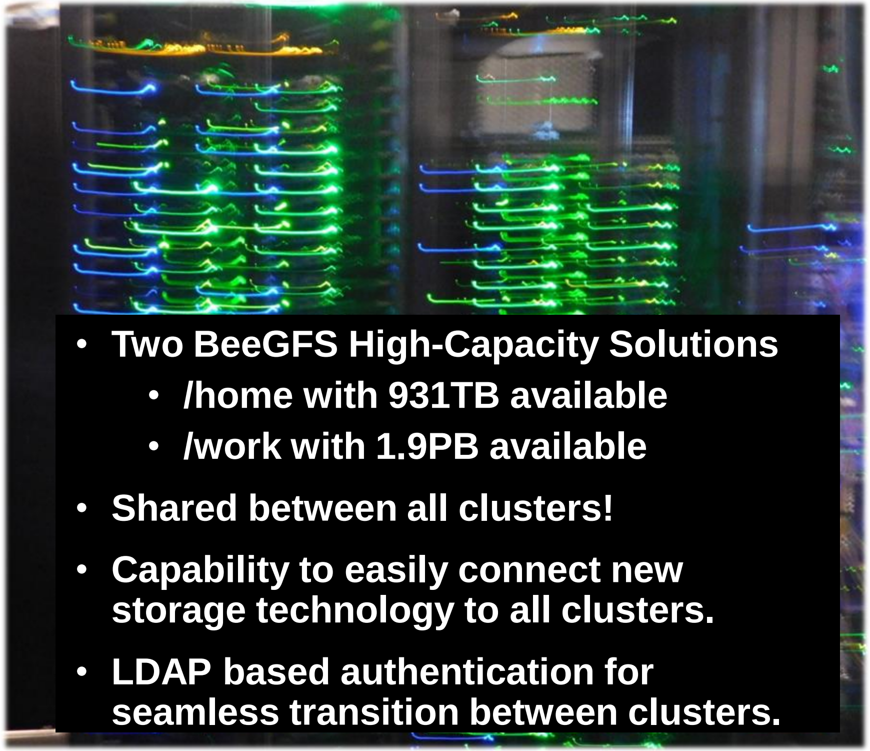
13K ft.^{2} lab, 1,300+ servers, ~10PB storage dedicated to HPC and AI in collaboration with the community

Zenith

- TOP500-class system
- Was #383, #292, #265, #396 on Top 500
- 420 servers Xeon servers, HDR100 InfiniBand
- **~1160 TF combined performance!**
- BeeGFS
- **Liquid** cooled and air cooled

Other systems

- Smaller test clusters, storage solutions, etc.



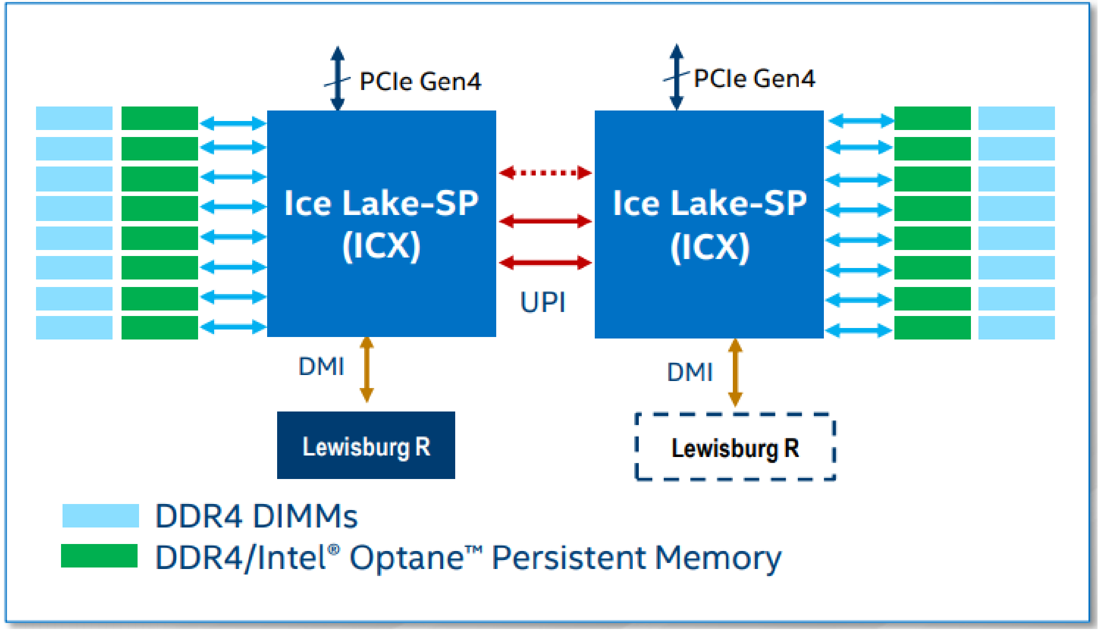
- **Two BeeGFS High-Capacity Solutions**
  - **/home with 931TB available**
  - **/work with 1.9PB available**
- **Shared between all clusters!**
- **Capability to easily connect new storage technology to all clusters.**
- **LDAP based authentication for seamless transition between clusters.**

Intel® Ice Lake Architecture

Intel Icelake 3rd Gen Intel Xeon Scalable Processors

- Faster UPI
- Faster I/O
- Enhanced memory performance
- Increased Memory capacity
- Intel Optane memory

Whitley 2-socket System



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

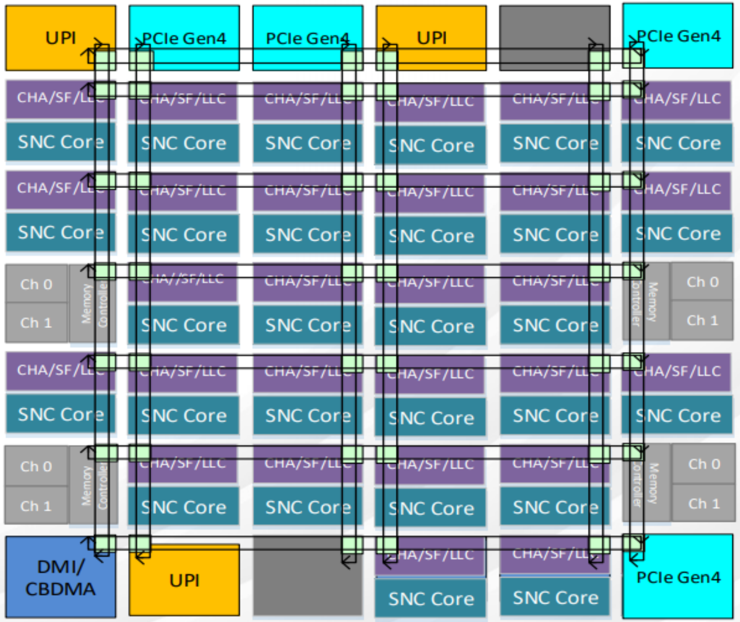
Intel® Total Memory Encryption (TME)

- DRAM encrypted using AES-XTS 128bit

Intel Optane Persistent Memory 200 Series (Barlow Pass)

- Higher speed and better power profile

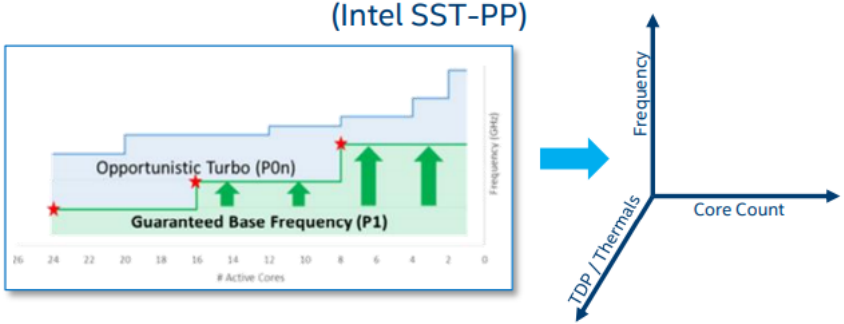
Ice Lake SP (28 core example)



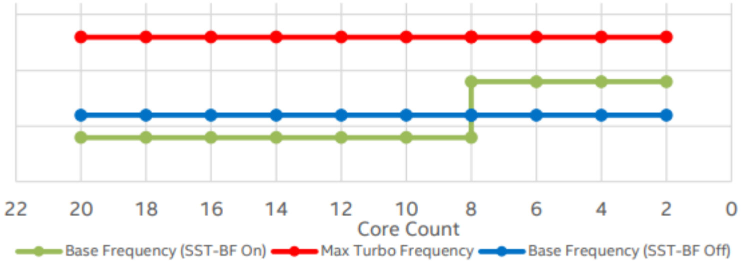
Intel® Speed Select Technology (Intel® SST) Features

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

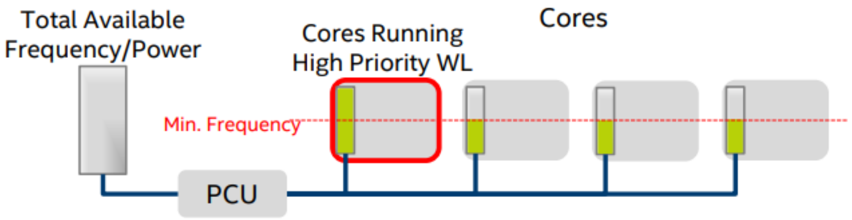
Intel Speed Select Technology–**Performance Profile** (Intel SST-PP)



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

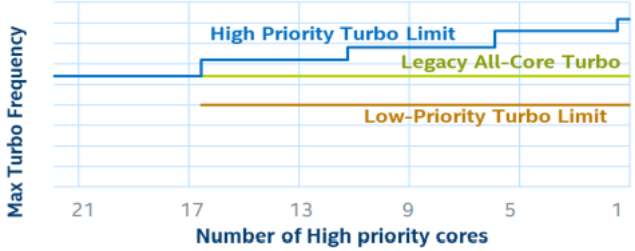


Intel Speed Select Technology–**Core Power** Intel SST-CP



Total Available Frequency/Power
Cores Running High Priority WL
Cores
Min. Frequency
PCU

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



The New 15th Generation Dell EMC PowerEdge Server Portfolio

Specialized– GPU OPTIMIZED

R750xa

R750

R650

C-SERIES

C6520

ALL NEW INTEL ICE LAKE
2-SOCKET SERVERS

YOUR INNOVATION ENGINE

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

Application performance and BIOS tuning for HPC

HPL (optimization and performance)

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

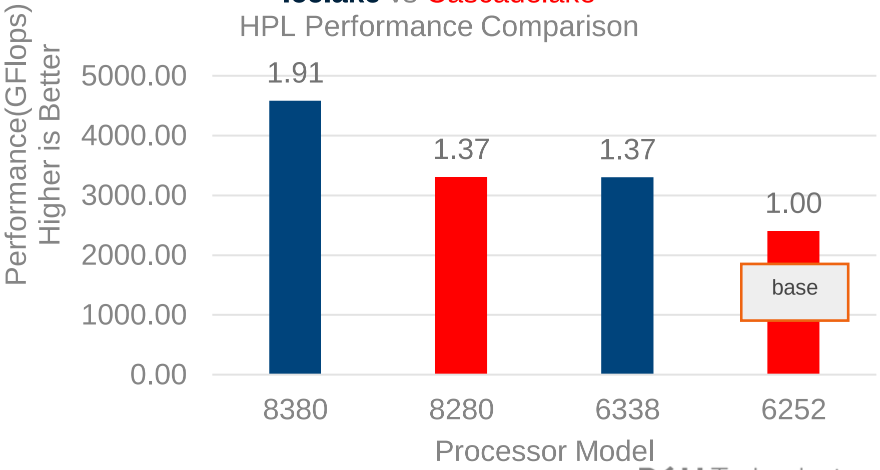
In case of open-source version of HPL -

- 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

Icelake vs Cascadelake
HPL Performance Comparison



STREAM Dual Socket(optimization and performance)

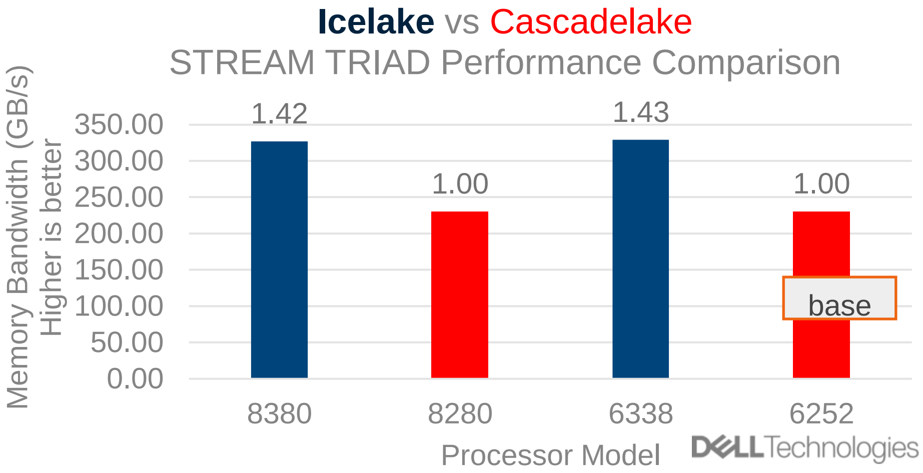
Intel compilers are recommended to get expected performance.

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

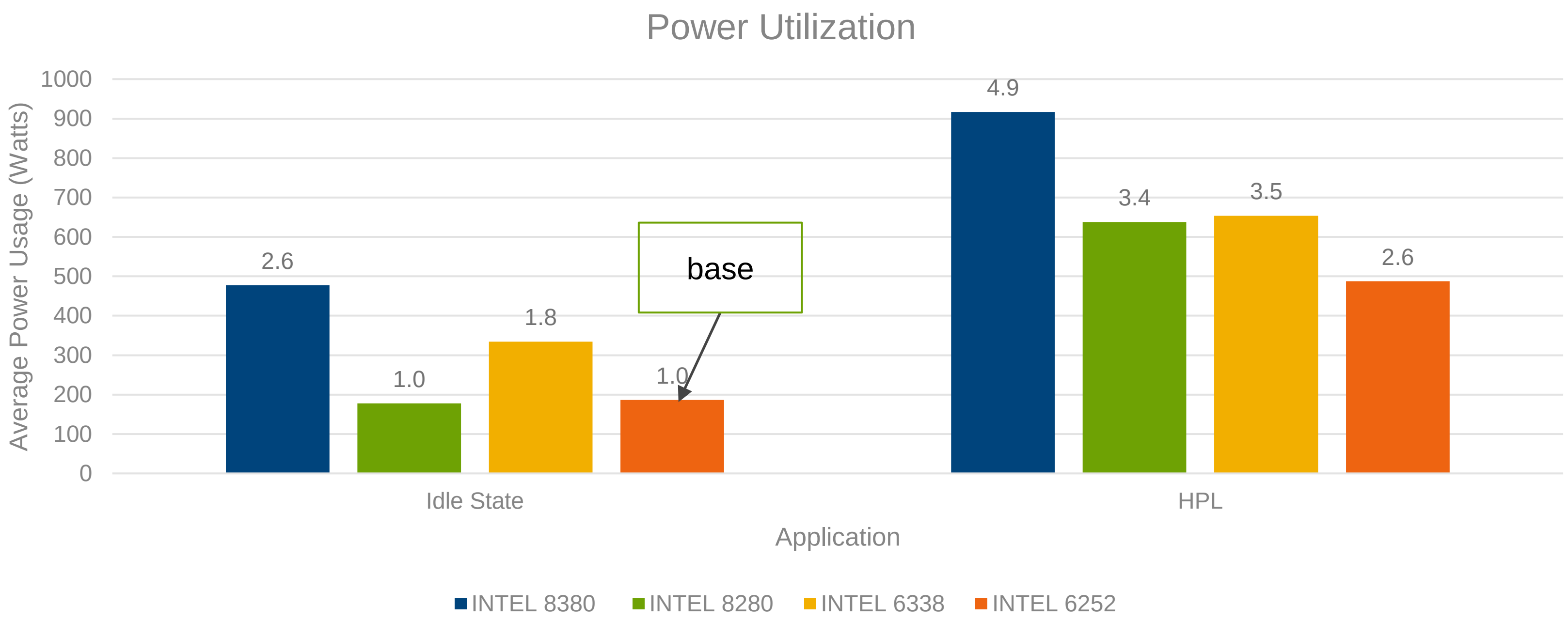
Recommended compiler flags (intel compiler) -
-xICELAKE-SERVER -O3 -ffreestanding -qopenmp -qopenmp-link=static -mcmodel=medium -shared-intel -restrict -qopt-streaming-stores **always** -DSTREAM_ARRAY_SIZE=160000000 -DNTIMES=100 -DOFFSET=0 -DVERBOSE -qopt-zmm-usage=high

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

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

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

Questions ?

intel®

DELL Technologies