Designing HPC Solutions

Onur Celebioglu

Dell Inc

Agenda

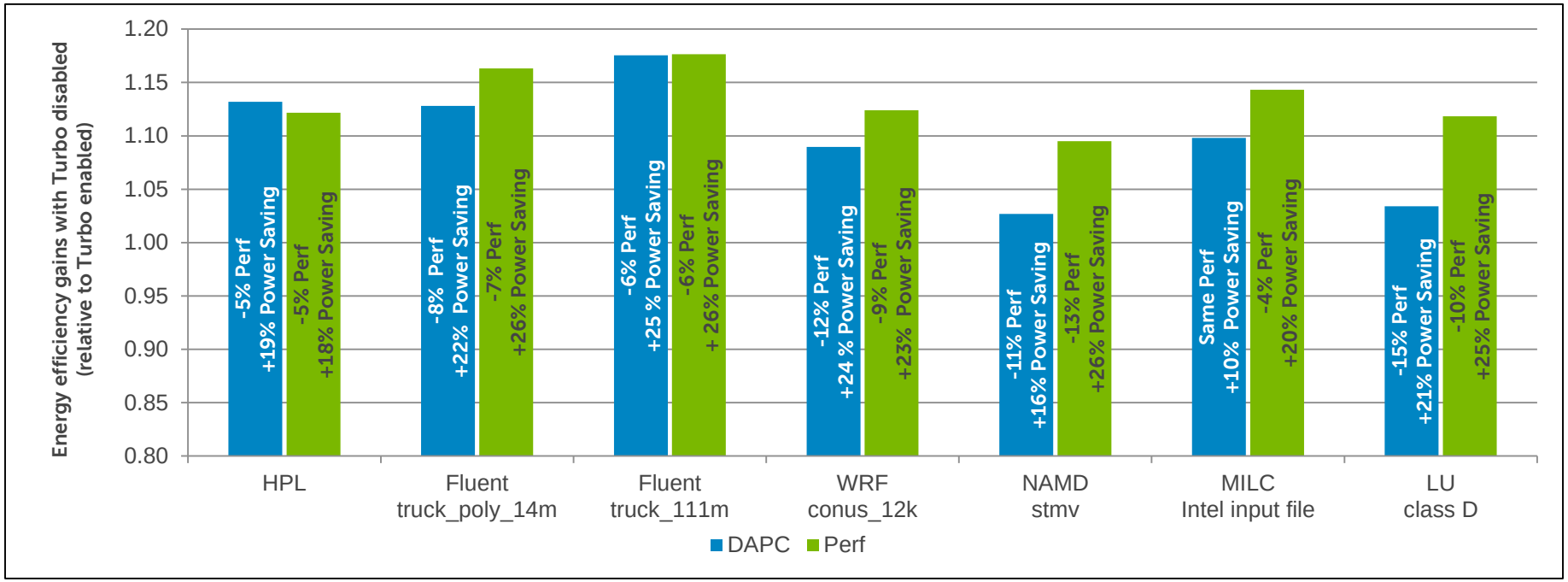
- HPC Focus Areas
- Performance analysis of HPC Components
  - Compute
  - Interconnect
  - Accelerators
  - And many more
- Best Practices
- Designing better HPC solutions
  - Domain specific Appliances

HPC at Dell

- Evaluate new HPC technologies and selectively adopt for Integration
- Share our findings with the broader HPC community.
- Analyze decision points to obtain the optimal solution to the problem at hand.
- Decision Points include but not limited to
  - Compute Performance
  - Memory Performance
  - Interconnect
  - Accelerators
  - Storage
  - Power / Energy Efficiency
  - Software Stack
  - Middleware
- Focus Areas
  - Define best practices by analyzing each and every component of an HPC cluster
  - Use these best practices to develop plug and play solutions targeted at specific HPC verticals such as Life sciences, Fluid Dynamics, High frequency trading etc.

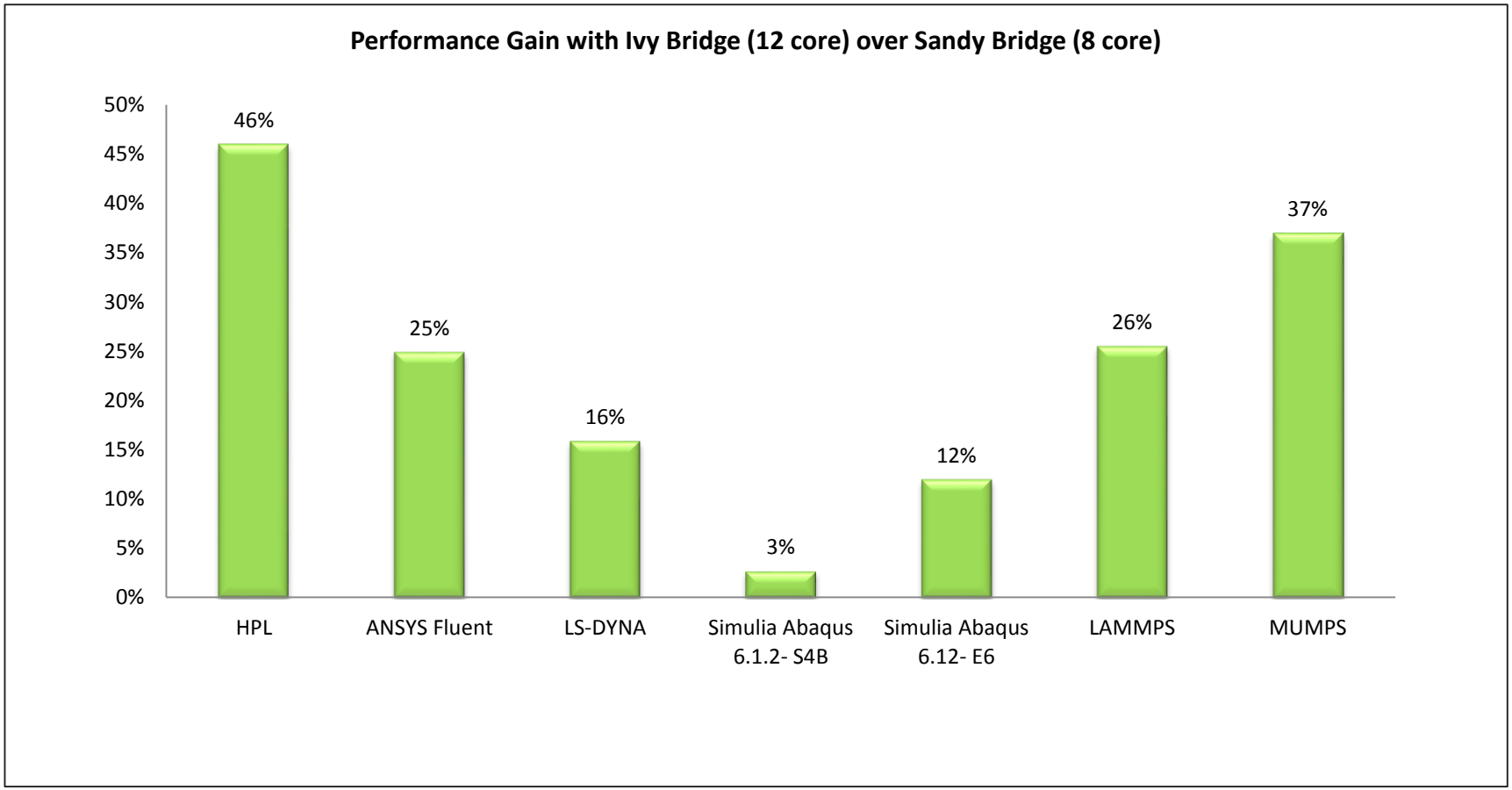
Compute, Memory & Energy Efficiency

12G – Optimal BIOS Settings



	Balanced configuration	Performance focused	Energy Efficient configuration	Latency sensitive
System Profile	Performance Per Watt Optimized (DAPC)	Performance Optimized	Custom	Custom
CPU Power Mgmt	System DBPM	Max Performance	System DBPM	Max Performance
Turbo Boost	Enabled	Enabled	Disabled	Disabled
C States & C1E	Enabled	Disabled	Enabled	Disabled
Monitor/ Mwait	Enabled	Enabled	Enabled	Disabled
Logical Processor	Disabled	Disabled	Disabled	Disabled
Node Interleaving	Disabled	Disabled	Disabled	Disabled

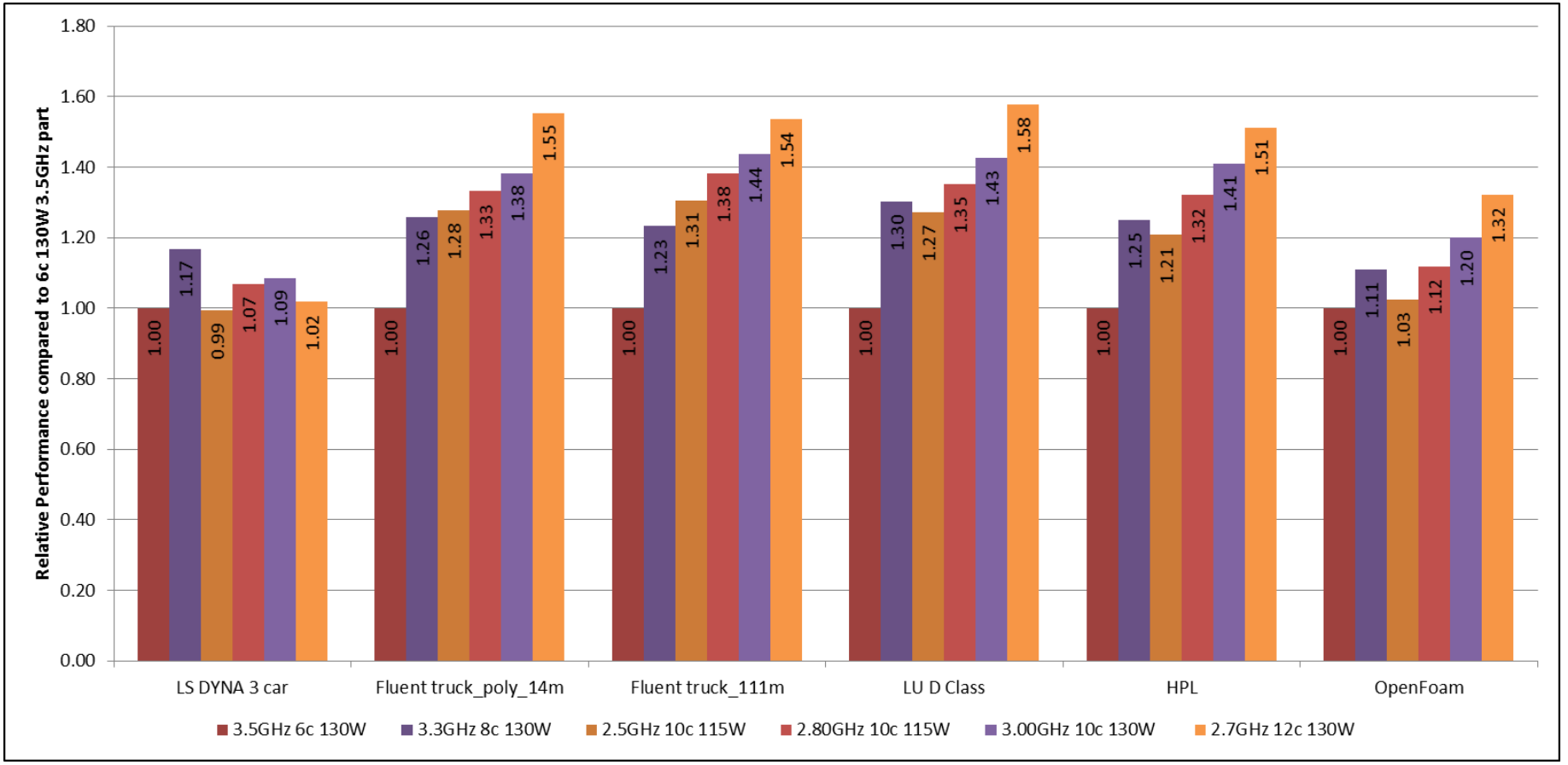
Ivy Bridge vs. Sandy Bridge Single Node



- E5-2670 8C 2.6 Ghz (SB) vs E5-2697 V2 12C 2.7 GHz (IVB)

Decision: Processor selection. Criteria: Performance

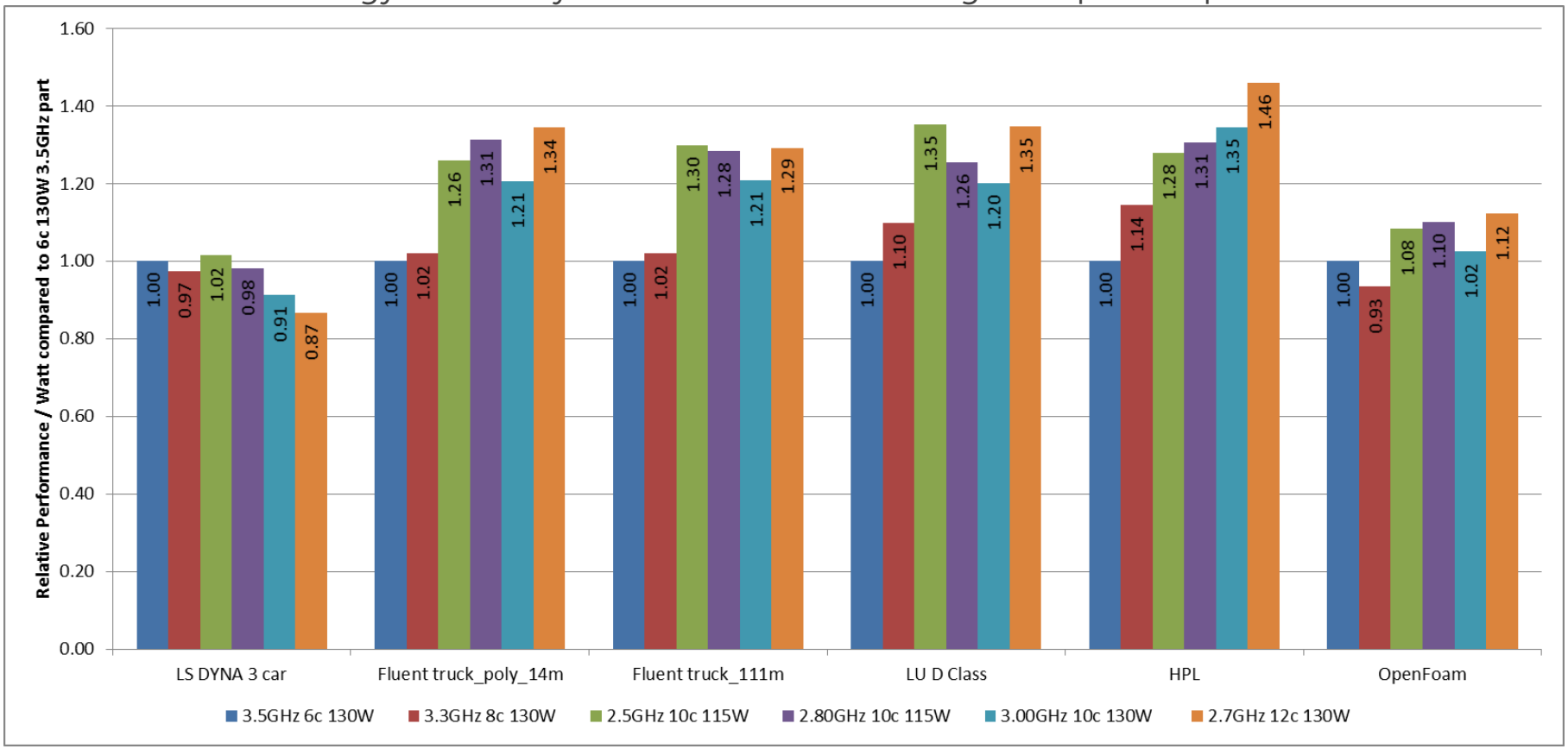
Performance across four nodes using multiple IVB processors



- 2 x E5-2697-v2 @ 2.7 GHz 12c 130W does the best in most cases.
- All tests done on fully subscribed 4 servers with FDR interconnect.

Decision: Processor selection. Criteria: Power

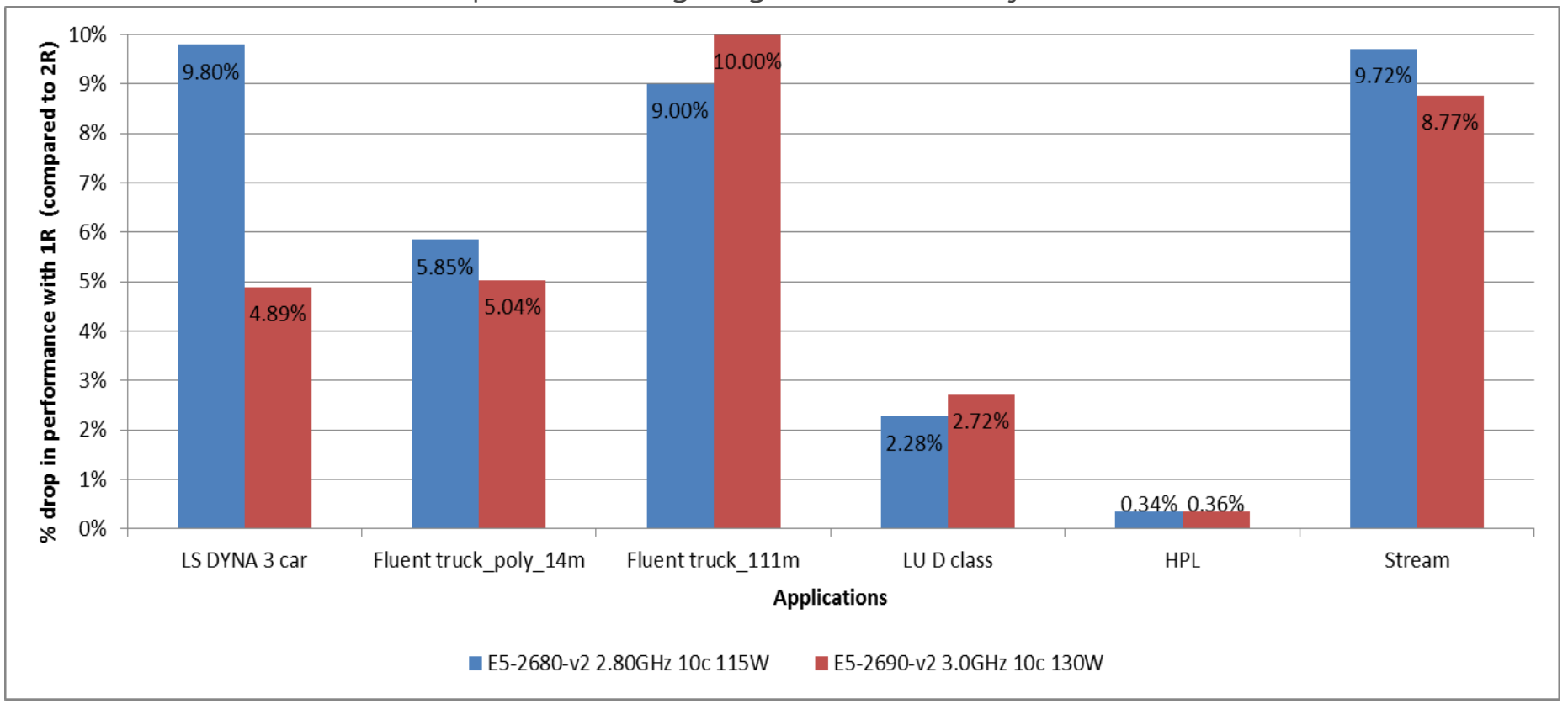
Energy efficiency across four nodes using multiple IVB processors



- 2 x E5-2697-v2 @ 2.7 GHz 12c 130W does the best in most cases.
- All tests done on fully subscribed 4 servers with FDR interconnect.

Decision: Memory selection. Criteria: Performance

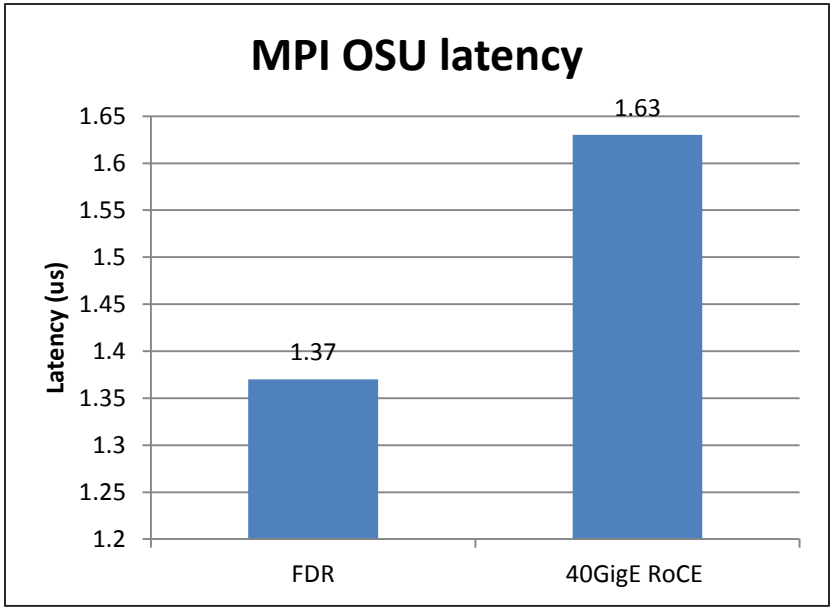
Performance drop when using single rank memory modules on 4 nodes

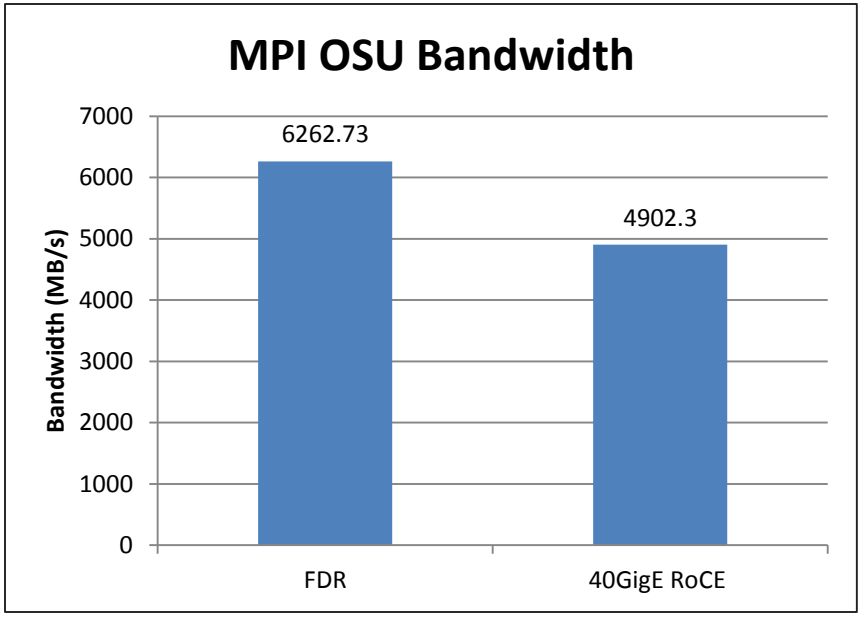


- Dual Rank memory modules give best performance
- All tests done on fully subscribed 4 servers with FDR interconnect.

Interconnect Performance

OSU Latency and Bandwidth (FDR vs 40 GigE RoCE)

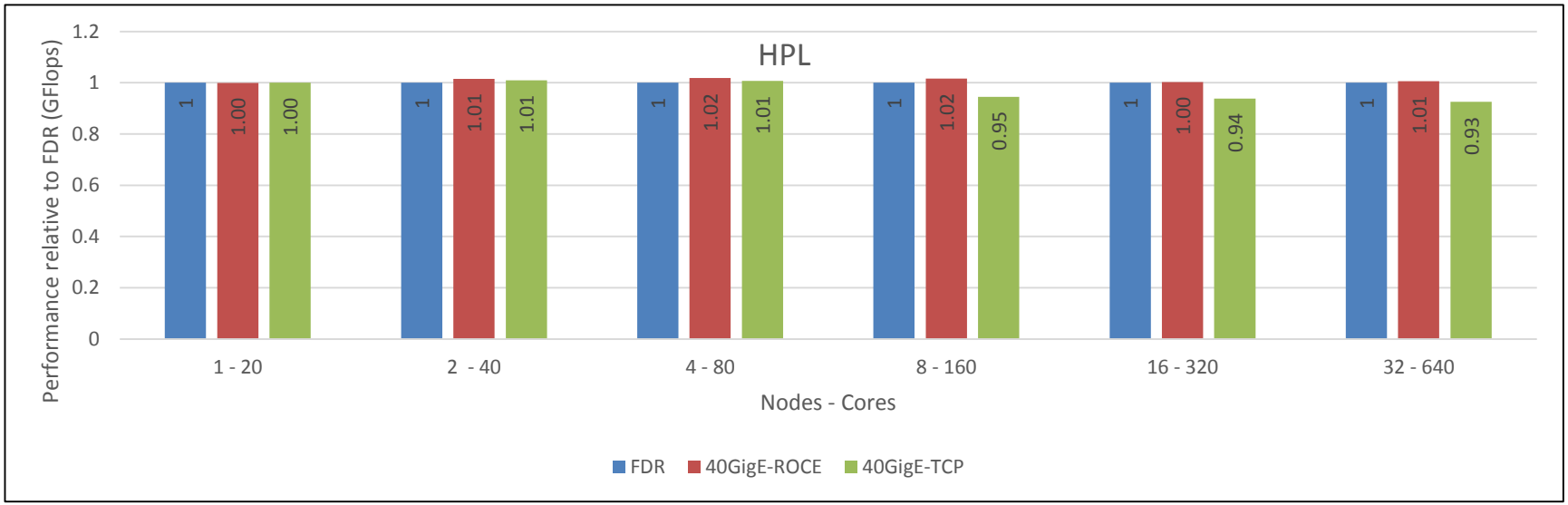


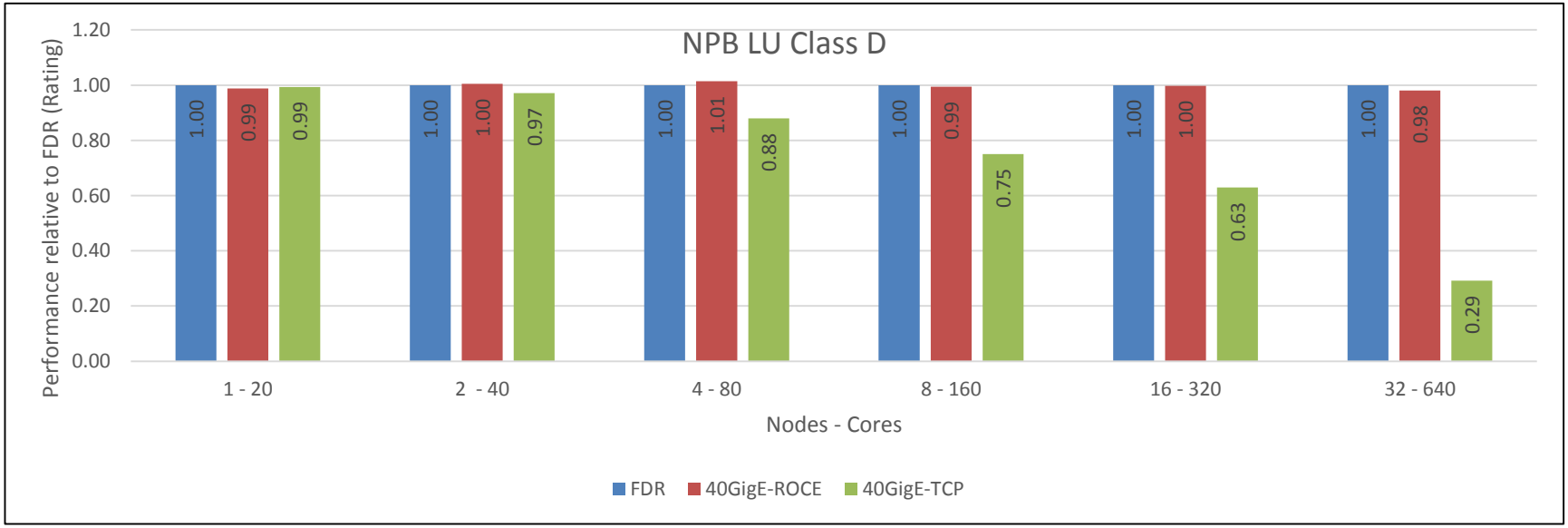


MVAPICH2-2.0b and OMB v4.2

- How do benchmarks, synthetic kernels and micro benchmarks behave at scale?
- Can micro benchmark performance explain application's performance at a larger scale?

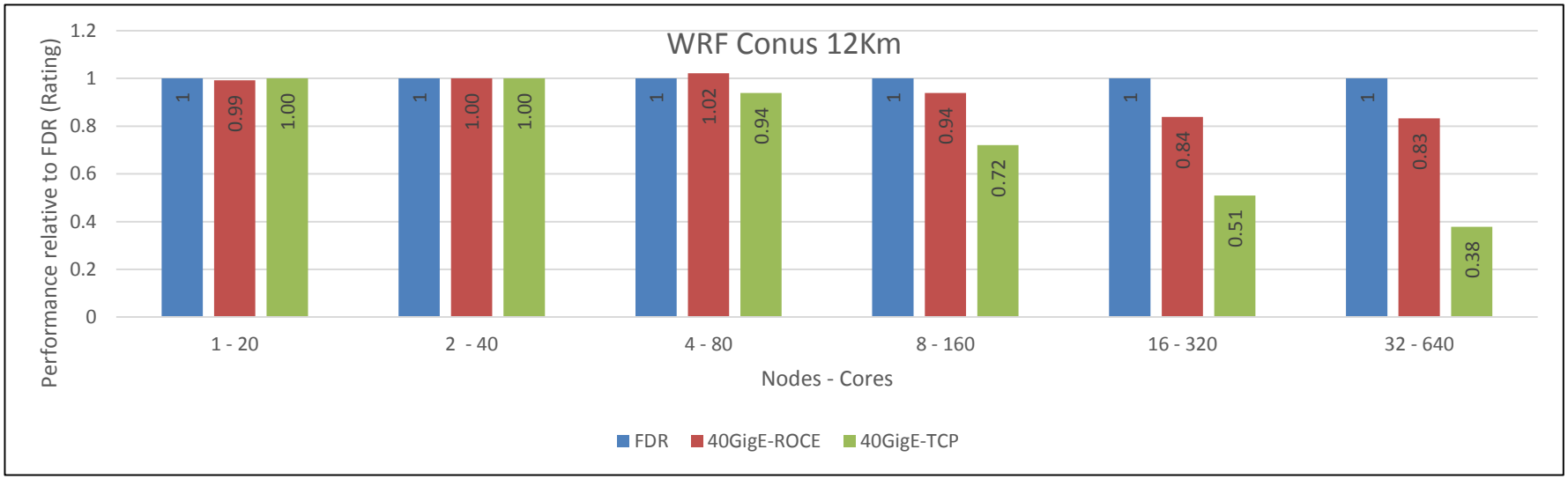
RoCE vs IB vs TCP

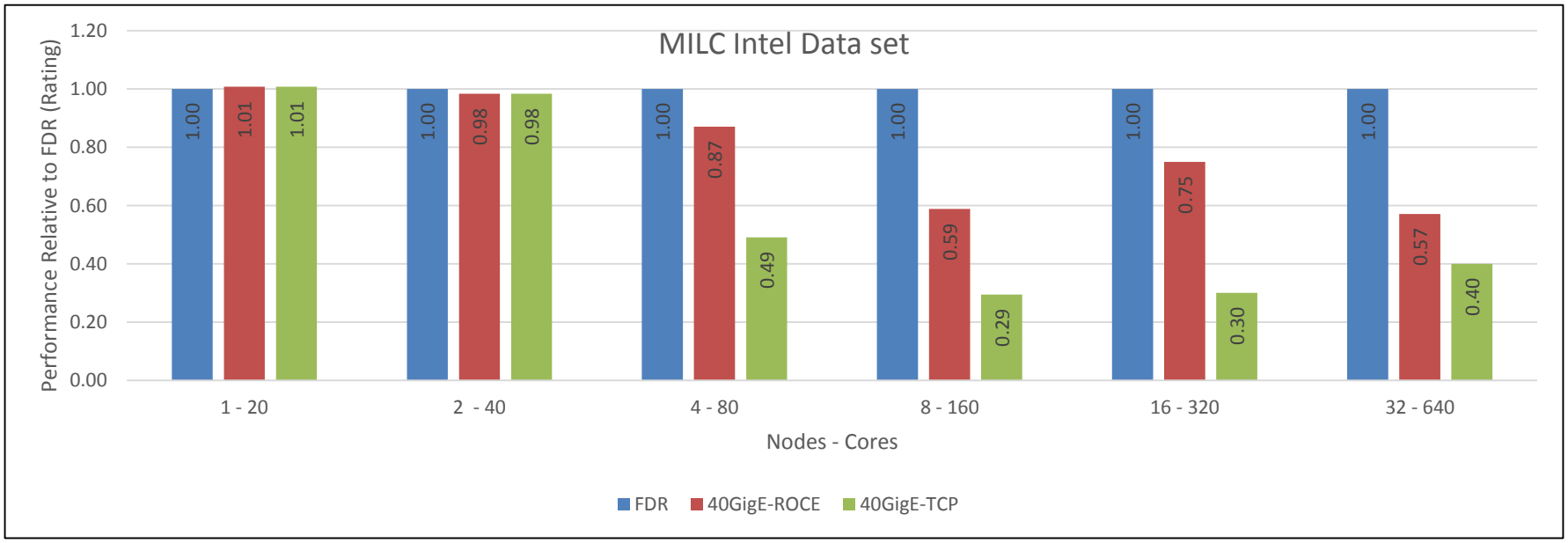




MVAPICH2-2.0b

RoCE vs IB vs TCP





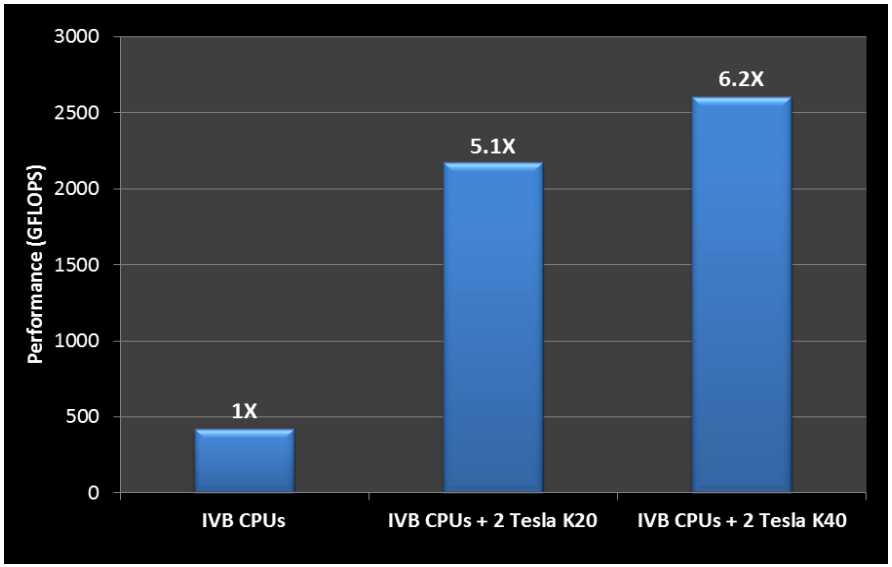
MVAPICH2-2.0b WRF 3.5 MILC 7.6.2

Interconnect Summary

- InfiniBand is still performs higher than other network fabrics in this study for HPC workloads
- For some workloads, RoCE performs similar to InfiniBand and may be a viable alternative.
  - Haven't seen wide adoption of RoCE in production yet.
  - Mileage will vary based on application's communication characteristics
  - Needs switches with DCB support for optimal lossless performance.
- Ethernet with TCP/IP stops scaling after 4-8 nodes.

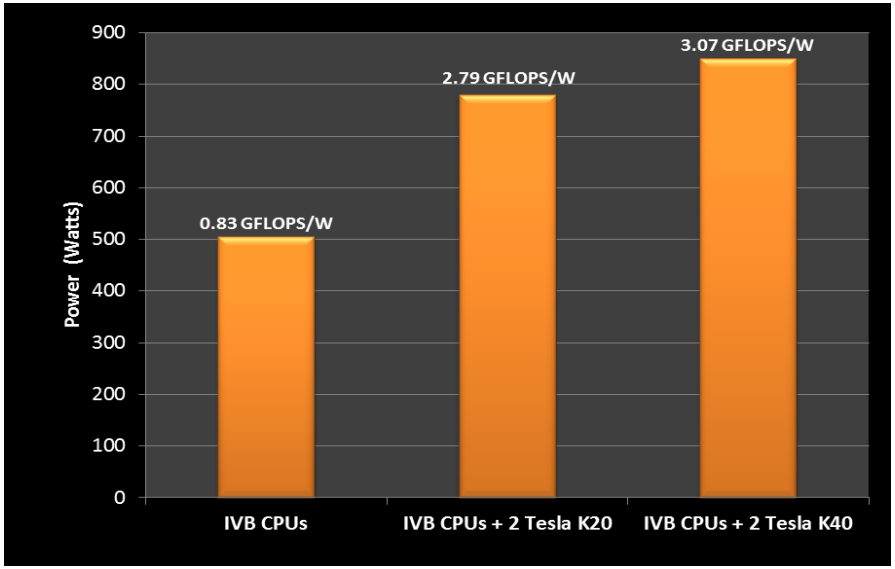
Accelerator Performance

Power and Performance: K20 vs K40



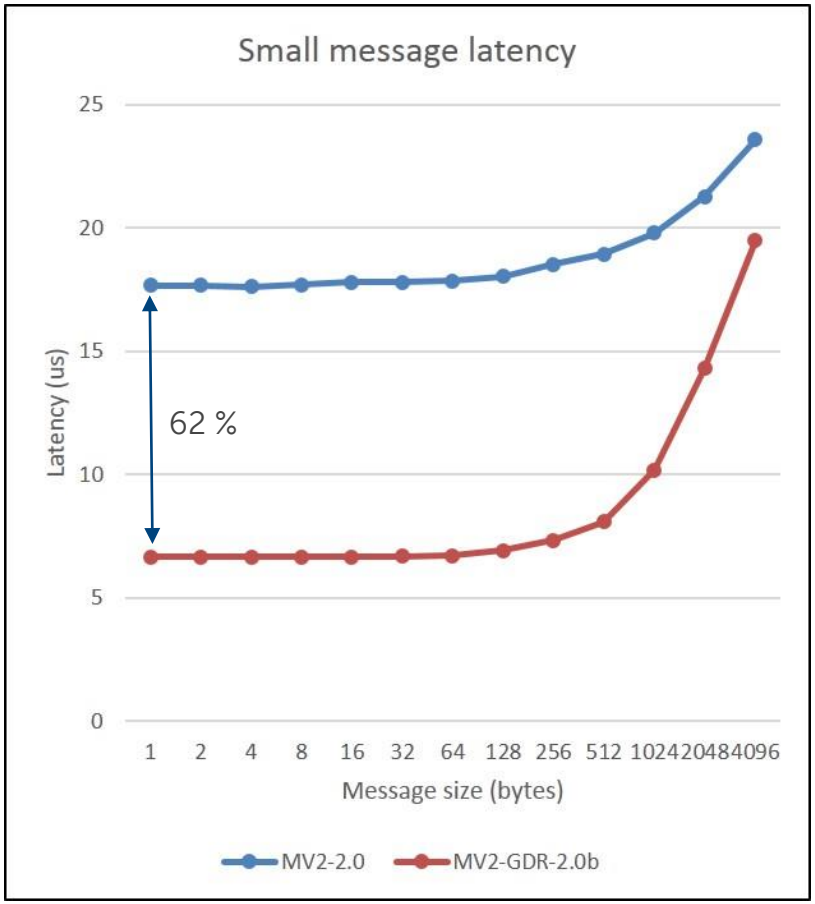
HPL performance on single-node

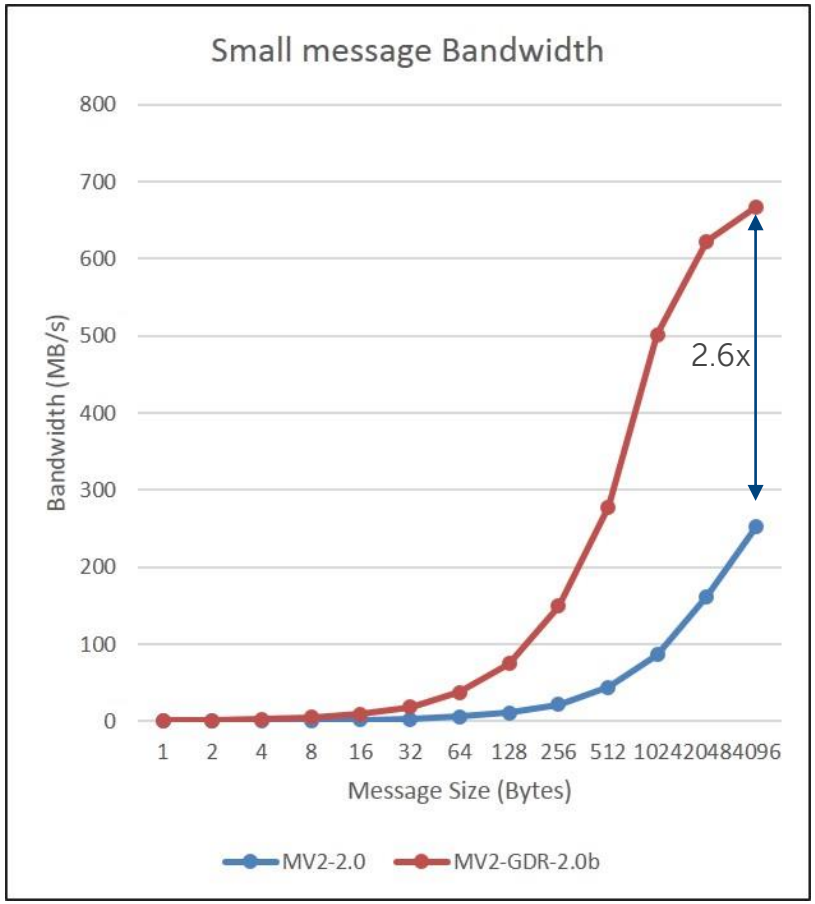
Power & energy efficiency of an eight node cluster.



MV2 performance with GPU Direct : OMB

Device to Device Latency





Intel Sandy Bridge (E5-2670) , NVIDIA Telsa K20m GPU, Mellanox ConnectX-3 FDR, CUDA 6.0, OFED 2.2-1.0.0 with GPU Direct RDMA Beta

Domain specific solutions

Dell Genomic Analysis Platform

Mellanox SX6036 IB Switch

Dell PowerEdge R320
HSS 4.5 Management Node
2 x Dell PowerEdge R420
BCM Login Nodes

Dell PowerEdge R820
Fat Node With 1.5 TB Memory

Dell PowerVault MD3260
NSS 4.5 HA Storage

Dell PowerVault MD3220
HSS 4.5 Lustre Metadata Storage Target

Dell PowerVault MD3260
HSS 4.5 Object Storage Target

Dell PowerEdge M1000e Chassis
With 32 Dell PowerEdge M420s

Force10 S55 Switch

Dell PowerEdge R320 CIFS Gateway

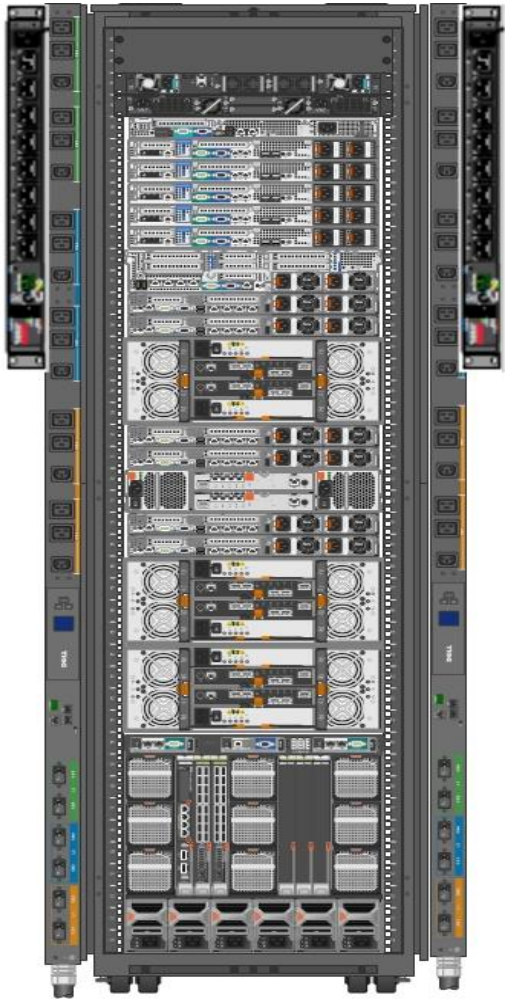
2 x Dell PowerEdge R420
BCM Head Nodes

2 x Dell PowerEdge R620
NSS 4.5 HA Head Nodes

2 x Dell PowerEdge R620
HSS 4.5 Lustre Metadata Servers

2 x Dell PowerEdge R620
HSS 4.5 Lustre Object Storage Servers

Dell PowerVault MD3260
HSS 4.5 Object Storage Target



Dell Genomic Analysis Platform (Continued)

Parameter	Results and Analysis
Time taken for analyzing 30 samples	19.5 Hours
Energy Consumption for analyzing 30 samples	222.77 kWh
kWh/Genome	7.42 kWh / Genome
Genomes/day	37

Advantages

- Metrics relevant to the domain instead of GFLOPs.
- Energy Efficient
- Plug and Play
- Scalability
- What used to take 2 weeks now takes less than 4 hours.
- More to follow..

Collateral

Future Work and Potential Areas of Research

- Deployment tools
- Use of virtualization and cloud (Openstack) in HPC
  - Linux Containers and Docker
- Hadoop
- Lustre FS
- Accelerators

Storage Blogs

- HTSS + DX Object Storage
  - http://dell.to/zJqiTK
- Dell HPC NFS Storage Solution with High Availability -- Large Capacity Configuration
  - http://dell.to/GYWU5x
- Dell support for XFS greater than 100 TB
  - http://dell.to/GUjXRq
- NSS-HA 12G Performance Intro
  - http://dell.to/NFUafG
- NSS4.5-HA Solution Configurations
  - http://dell.to/10xLxJV
- Dell Fluid Cache for DAS performance with NFS
  - http://dell.to/15KnsDc
- Achieving over 100000 IOPs with NFS Async
  - http://dell.to/16yE3bP
- Dell | Terascala HPC Storage Solution Part I
  - http://www.delltechcenter.com/page/Dell+|+Terascala+HPC+Storage+Solution+Part+I
- Dell | Terascala HPC Storage Solution Part 2
  - http://en.community.dell.com/techcenter/high-performance-computing/w/wiki/2336.aspx
- DT-HSS3 Performance and Scalability
  - http://en.community.dell.com/techcenter/high-performance-computing/w/wiki/2300.aspx

Storage Blogs Continued

- Dell | Terascala HPC Storage Solution - HSS5
  - http://dell.to/1gpVVyN
- NSS overview
  - http://en.community.dell.com/techcenter/high-performance-computing/w/wiki/2338.aspx
- NSS-HA overview
  - http://en.community.dell.com/techcenter/high-performance-computing/w/wiki/2298.aspx
- NSS-HA XL configuration
  - http://en.community.dell.com/techcenter/high-performance-computing/w/wiki/2299.aspx
- Dell HPC NFS Storage Solution - High Availability Solution NSS5-HA configurations
  - http://dell.to/1eZU0xL

Coprocessor Acceleration Blogs

- GPUDirect Improves Communication Bandwidth Between GPUs on the C410X
  - http://dell.to/ApnLz5
- Comparing GPU-Direct Enabled Communication Patterns for Oil and Gas Simulations
  - http://dell.to/JsWqWT
- Accelerating ANSYS Mechanical Simulations with M2090 GPU on the R720
  - http://dell.to/JT79KF
- Accelerating High Performance Linpack (HPL) with GPUs
  - http://dell.to/MrYw8q
- Faster Molecular Dynamics with GPUs
  - http://dell.to/PEaFaF
- Deploying and Configuring Intel Xeon Phi Coprocessor with HPC Solution
  - http://dell.to/14GtFRv

Best Practices Blogs

- 12G HPC Solution with ROCKS+ from StackIQ
  - http://dell.to/xGmSHO
- HPC mode on Dell PowerEdge R815 with AMD 6200 Processors
  - http://dell.to/MMGG4s
- Optimal BIOS settings for HPC workloads
  - http://dell.to/PkkMG1
- CFD Primer
  - http://dell.to/UwJQum
- OpenFOAM
  - http://dell.to/Rga3hS
- PowerEdge M420 with single Force10 MXL Switch
  - http://dell.to/Zjnhjz
- Active Infrastructure for HPC Life Sciences
  - http://dell.to/18eaDSJ
- Dell HPC Solution Refresh: Intel Xeon Ivy Bridge-EP, 1866 DDR3 memory and RHEL 6.4
  - http://dell.to/18U3Aki

Performance Blogs

- HPC I/O performance using PCI-E Gen3 slots on the 12th Generation (12G) PowerEdge Servers
  - http://dell.to/wzdV0x
- HPC performance on the 12th Generation (12G) PowerEdge Servers
  - http://dell.to/zozohn
- Unbalanced Memory Configuration Performance
  - http://dell.to/UQ1kQu
- Performance analysis of HPC workloads
  - http://dell.to/STbE8q

Questions?