



Gateways to Discovery: Cyberinfrastructure for the Long Tail of Science

ECSS Symposium, 12/16/14

**M. L. Norman, R. L. Moore, D. Baxter, G. Fox (Indiana U), A
Majumdar, P Papadopoulos, W Pfeiffer, R. S. Sinkovits, S. Strande
(NCAR), M. Tatineni, R. P. Wagner, N. Wilkins-Diehr, UCSD/SDSC**



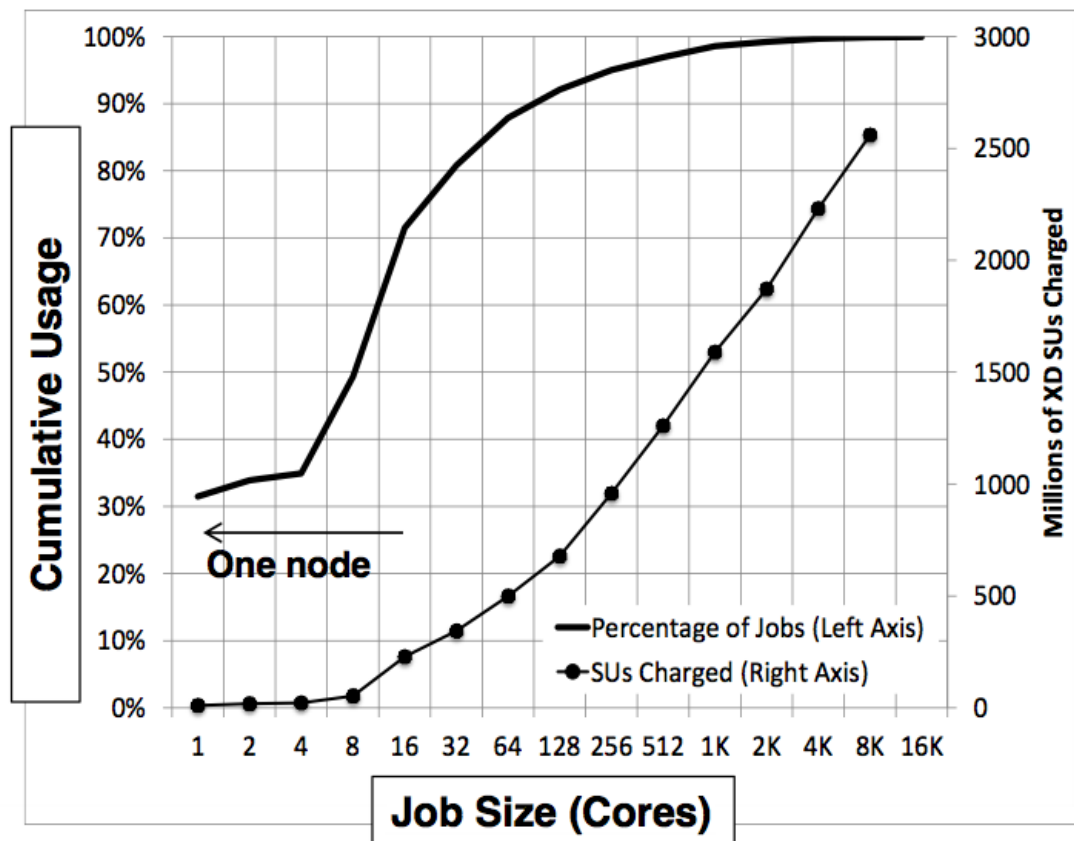
HPC for the 99%

High-performance computing for the long tail of science

- **Comet goals (from NSF 13-528 solicitation)**
 - “... expand the use of high end resources to a much larger and more diverse community
 - ... support the entire spectrum of NSF communities
 - ... promote a more comprehensive and balanced portfolio
 - ... include research communities that are not users of traditional HPC systems.”

HPC for the 99%

- 99% of jobs run on NSF's HPC resources in 2012 used <2,048 cores
- And consumed >50% of the total core-hours across NSF resources



Key Strategies for Comet Users

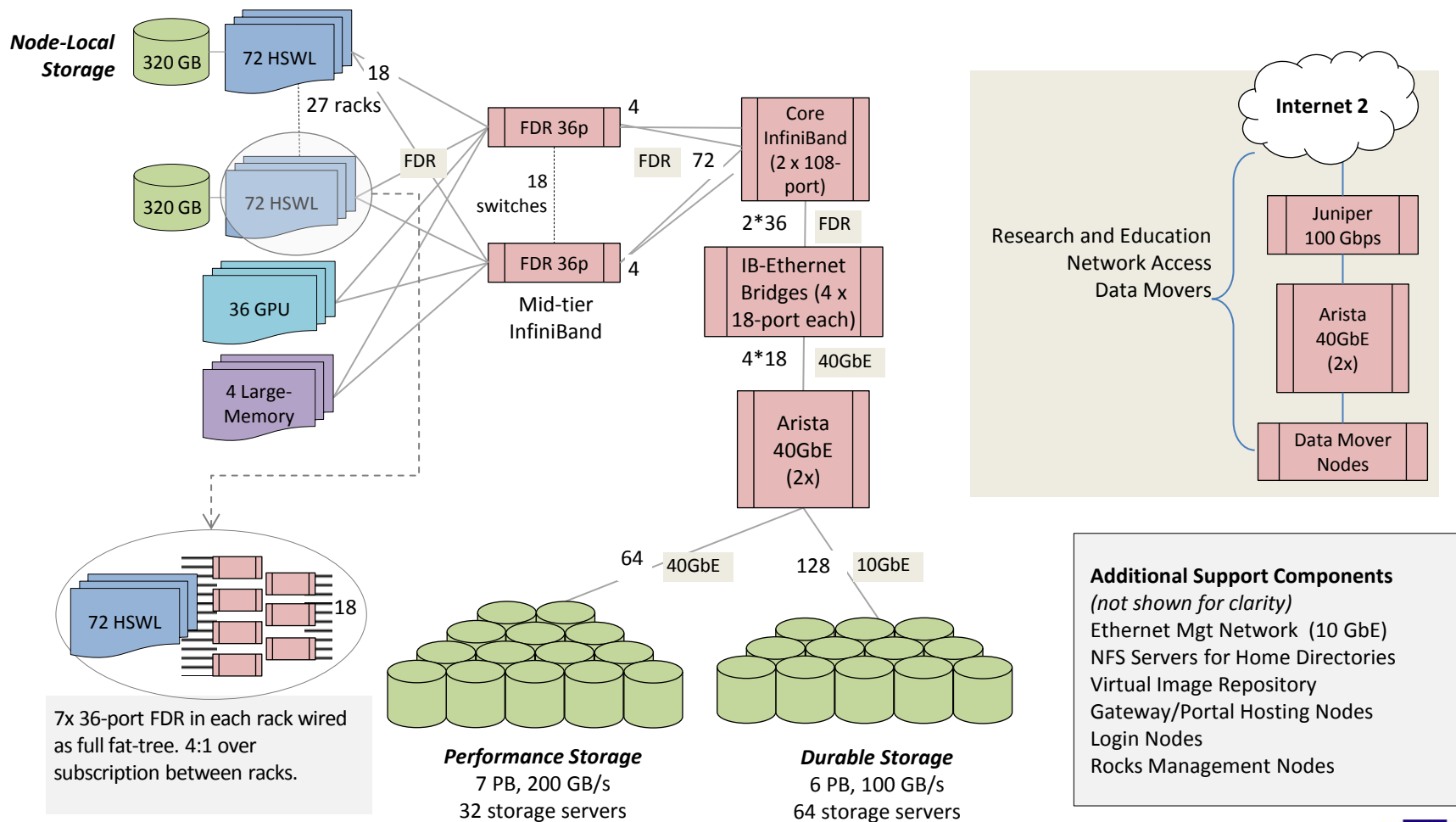
- Target modest-scale users and new users/communities: goal of 10,000 users/year!
- Support capacity computing, with a system optimized for small/modest-scale jobs and quicker resource response using allocation/scheduling policies
- Build upon and expand efforts with Science Gateways, encouraging gateway usage and hosting via software and operating policies
- Provide a virtualized environment to support development of customized software stacks, virtual environments, and project control of workspaces

Comet: System Characteristics

- **Total peak flops 2 PF**
- **Dell primary integrator**
 - Intel Haswell processors w/ AVX2
 - Mellanox FDR InfiniBand
- **1,944 standard compute nodes (47K cores)**
 - Dual CPUs, each 12-core, 2.5 GHz
 - 128 GB DDR4 2133 MHz DRAM
 - 2*160GB GB SSDs (local disk)
- **36 GPU nodes (Feb 2015)**
 - Same as standard nodes *plus*
 - Two NVIDIA K80 cards, each with dual Kepler3 GPUs
- **4 large-memory nodes (April 2015)**
 - 1.5 TB DDR4 1866 MHz DRAM
 - Four Haswell processors/node
- **Hybrid fat-tree topology**
 - FDR (56 Gbps) InfiniBand
 - Rack-level (72 nodes, 1,728 cores) full bisection bandwidth
 - 4:1 oversubscription cross-rack
- **Performance Storage (Aeon)**
 - 7.6 PB, 200 GB/s; Lustre
 - Scratch & Persistent Storage segments
- **Durable Storage (Aeon)**
 - 6 PB, 100 GB/s; Lustre
 - Automatic backups of critical data
- **Gateway hosting nodes**
- **Virtual image repository**
- **Home directory storage**
- **100 Gbps external connectivity to Internet2 & ESNet**

Comet Network Architecture

InfiniBand compute, Ethernet Storage



Suggested Comet Applications

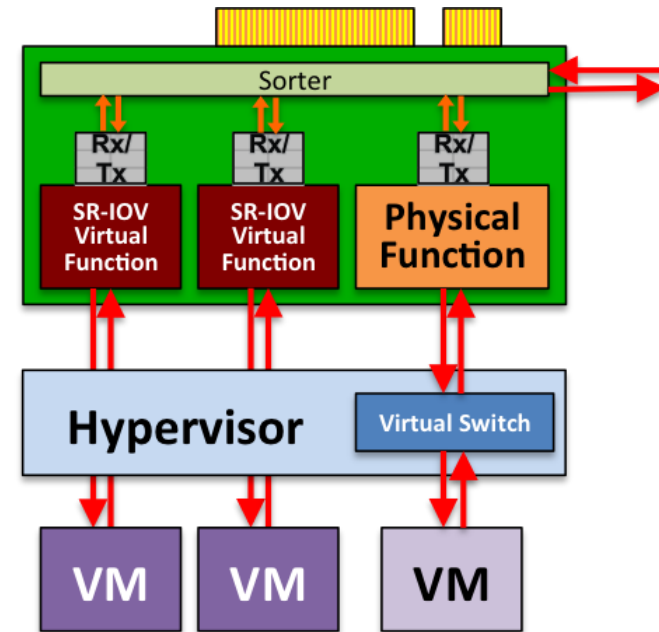
- **Modest core counts:** full bisection bandwidth up to Comet island (1,728 cores)
- **128 GB DRAM/node (5.3 GB/core):** single node shared memory apps and MPI codes with large per-process memory footprint
- **AVX2:** Codes with vectorizable loops. Any application with significant performance gain relative to Sandy Bridge or Ivy Bridge (AVX)
- **SSDs:** Computational chemistry, finite elements. Apps that generate large numbers of small temporary files (finance, QM/MM)

Suggested Comet Applications, cont'd

- **GPU nodes:** Molecular dynamics, linear algebra, image and signal processing.
 - Doesn't replace Keeneland, but for workloads that have some GPU requirements.
- **Large memory nodes:** *de novo* genome assembly, visualization of large data sets, other large memory apps
- **Science Gateways:** Gateway-friendly environment with local gateway hosting capability, flexible allocations, scheduling policies for rapid throughput, heterogeneous workflows, and virtual clusters for software environment
- **High performance virtualization:** workloads with customized software stacks, especially those that are difficult to port or deploy in standard XSEDE environment

Single Root I/O Virtualization in HPC

- **Problem:** Virtualization generally has resulted in significant I/O performance degradation (e.g., excessive DMA interrupts)
- **Solution:** SR-IOV and Mellanox ConnectX-3 InfiniBand host channel adapters
 - One physical function → multiple virtual functions, each light weight but with its own DMA streams, memory space, interrupts
 - Allows DMA to bypass hypervisor to VMs
- ***SRIOV enables virtual HPC cluster w/ near-native InfiniBand latency/bandwidth and minimal overhead***



Latency Results:

QDR IB & 10 GbE, native and SR-IOV

- SR-IOV with QDR InfiniBand
 - < 30% overhead for small messages (<128 bytes)
 - < 10% overhead for eager send/receive
 - Overhead → 0% for bandwidth-limited regime
- Amazon EC2 (10 GbE)
 - > 50X worse latency
 - Time dependent (noisy)

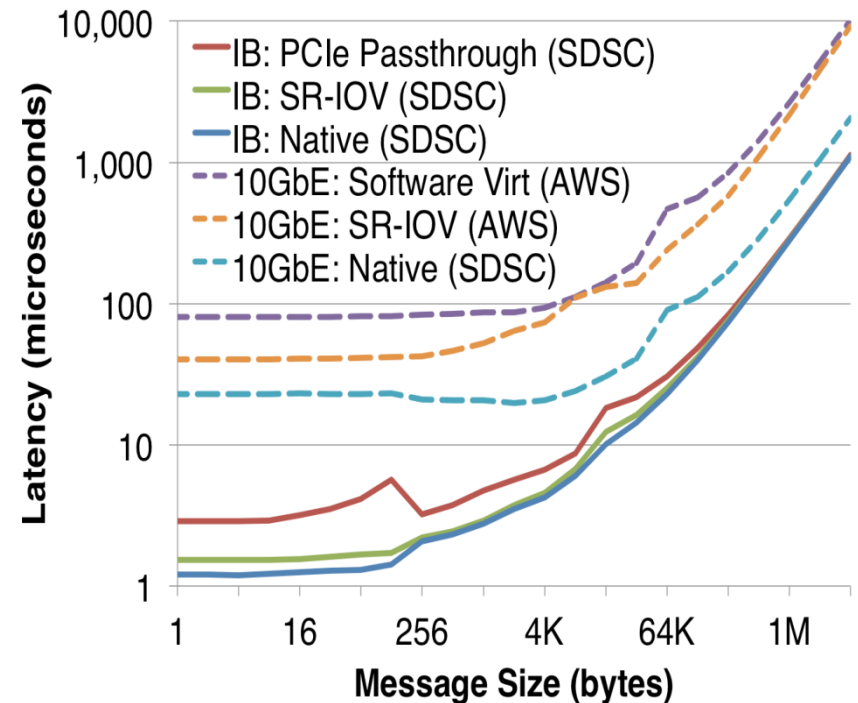


Figure 5. MPI point-to-point latency measured by `osu_latency` for QDR InfiniBand. Included for scale are the analogous 10GbE measurements from Amazon (AWS) and non-virtualized 10GbE.

50x less latency than Amazon EC2

Bandwidth Results: QDR IB & 10 GbE, native and SR-IOV

- Comparison of bandwidth relative to native InfiniBand
- SR-IOV w/ QDR InfiniBand
 - < 2% bandwidth loss over entire range
 - > 95% peak bandwidth
- Amazon EC2 (10 GbE)
 - < 35% peak bandwidth
 - While ratio of QDR/10GbE bandwidth is ~4X, EC2 bandwidth is 9-25X worse than SR-IOV IB

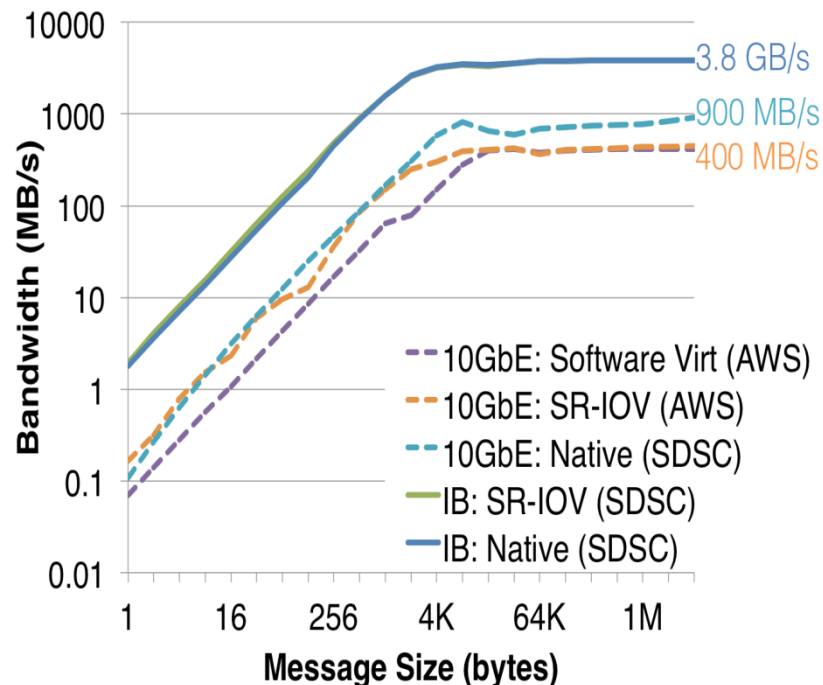
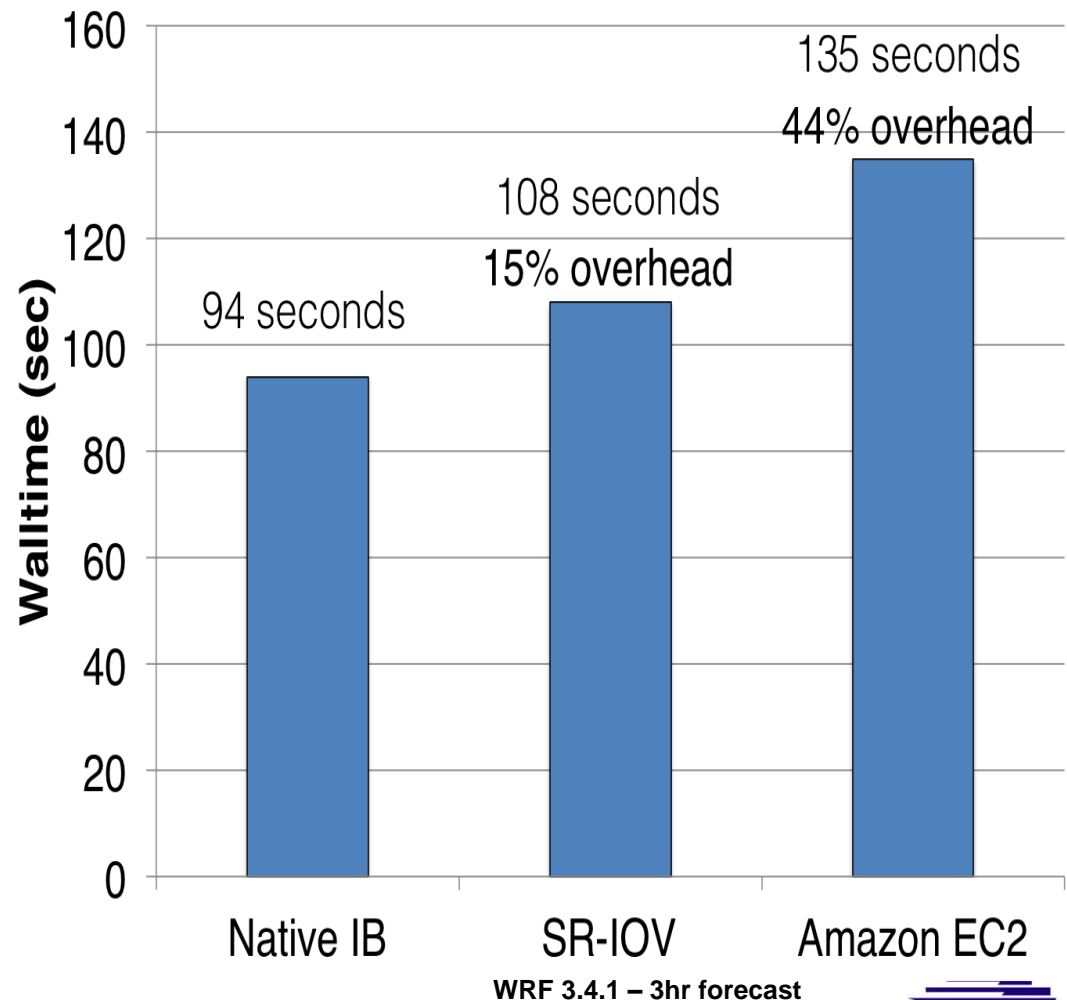


Figure 6. MPI point-to-point bandwidth measured by `osu_bw` for QDR InfiniBand. Included for scale are the analogous 10GbE measurements from Amazon (AWS) and non-virtualized 10GbE.

10x more bandwidth than Amazon EC2

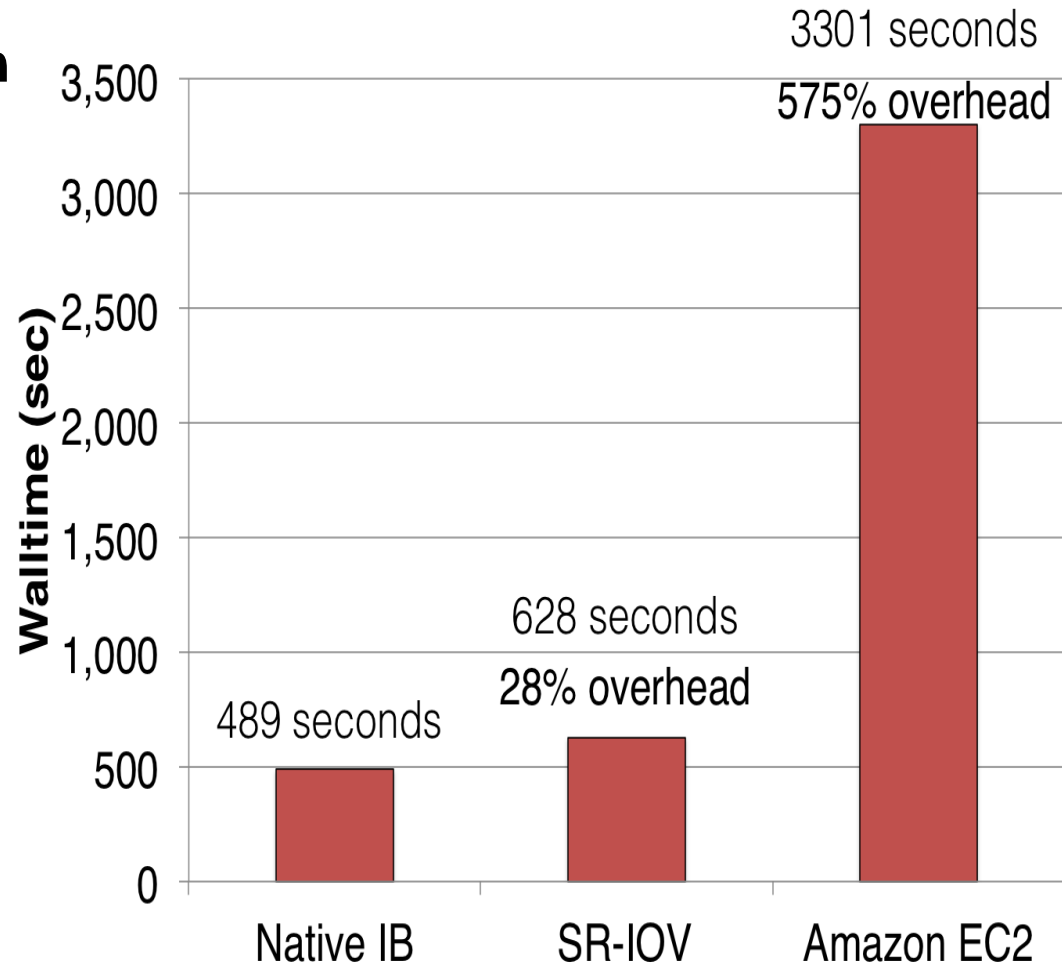
WRF Weather Modeling – 15% Overhead with SR-IOV IB

- **96-core (6-node) calculation**
- **Nearest-neighbor communication**
- **Scalable algorithms**
- **SR-IOV incurs modest (15%) performance hit**
- **... but still 20% faster than EC2**
 - Despite 20% slower CPUs



Quantum ESPRESSO: 28% Overhead

- 48-core (3 node) calculation
- CG matrix inversion - irregular communication
- 3D FFT matrix transposes (all-to-all communication)
- 28% slower w/ SR-IOV vs native IB
- SR-IOV still > 500% faster than EC2
 - Despite 20% slower CPUs



Quantum Espresso 5.0.2 – DEISA AUSURF112 benchmark

High Level Schedule

- **Dec 2014-Jan 2015**
 - Build and component test
- **Feb 2015**
 - Friendly users
 - Integrated acceptance tests
 - NSF review panel
- **March 2015**
 - production



UC San Diego



SDSC
SAN DIEGO SUPERCOMPUTER CENTER



INDIANA UNIVERSITY



This work supported by the National Science Foundation, award ACI-1341698.

SDSC SAN DIEGO SUPERCOMPUTER CENTER

at the UNIVERSITY OF CALIFORNIA; SAN DIEGO

