



# Intelligent Networking with SDN: Current State of the Art, Emerging Technologies and Future Visions

Panel at IONI '14

by

**Dhabaleswar K. (DK) Panda**

The Ohio State University

E-mail: [panda@cse.ohio-state.edu](mailto:panda@cse.ohio-state.edu)

<http://www.cse.ohio-state.edu/~panda>



# Major Systems in Geographically Distributed Environment

- Labs and Campuses with following systems
  - Supercomputers
  - Data Centers
  - Storage Area Networks
  - Data Transfer Nodes (DTNs)
- Connected with ESNNet and Internet2

# Current Trends in Networking Technology in Supercomputers, Parallel File Systems and Data Centers

- **Wide adaptation of InfiniBand technology with Remote Direct Memory Access (RDMA) mechanism**
  - 1.0 micro-sec node-to-node latency and 100Gbps unidirectional bandwidth (MPI-Level) with Mellanox Connect-IB dual-port adapters
- **223 Systems (44.3%) in June 2014 TOP500 List**
  - 25 systems in the TOP 50
- **Other systems (Cray, K-supercomputer, IBM Blue Gene) also use RDMA**
- **Convergence of InfiniBand and Ethernet -> RDMA over Converged Enhanced Ethernet (RoCE)**
  - Available in 10Gbps and 40Gbps
- **Datacenters are moving to InfiniBand/RoCE**
- **WAN is also adapting RDMA technology**

# Towards Exascale System (Today and Target)

Systems	2014 Tianhe-2	2020-2022	Difference Today & Exascale
System peak	55 PFlop/s	1 EFlop/s	~20x
Power	18 MW (3 Gflops/W)	~20 MW (50 Gflops/W)	O(1) ~15x
System memory	1.4 PB (1.024 PB CPU + 0.384 PB CoP)	32 – 64 PB	~50X
Node performance	3.43TF/s (0.4 TF/s CPU + 3 TF/s CoP)	1.2 or 15 TF	O(1)
Node concurrency	24 core CPU + 171 cores CoP	O(1k) or O(10k)	~5x - ~50x
Total node interconnect BW	6.36 GB/s	200 – 400 GB/s	~40x -~60x
System size (nodes)	16,000	O(100,000) or O(1M)	~6x - ~60x
Total concurrency	3.12M 12.48M threads (4 /core)	O(billion) for latency hiding	~100x
MTTI	Few/day	Many/day	O(?)

Courtesy: Prof. Jack Dongarra

# Can SDN be brought inside Supercomputers for E2E Connectivity?

- SDN reconfiguration has overhead
- Depends on applications and communication patterns of HPC and Big Data applications
  - OK if applications have regular communication patterns and large messages
  - Many applications have small to medium messages and irregular communication patterns
- Mostly No

# How about Parallel File Systems, Storage Area Networks and DTNs?

- Parallel File Systems and Storage Area Networks
  - Dual roles
  - Data in/out from supercomputers
  - Data in/out from WAN through DTNs
- Has significant potential to exploit SDN capabilities together with new designs for DTNs
- Has impact on designing new and intelligent
  - Network services
  - Middleware
  - E2E application services

# Examples and Challenges

- Designing SDN-aware Parallel File Systems, Storage Area Networks and DTNs
  - Combining RDMA with SDN capabilities for Bulk Data Movement
  - Unified memory sub-system design between WAN transfer and file I/O
  - Efficient key-value storage for handling metadata
- Network Services
  - E2E bandwidth allocation between local and remote file systems
  - E2E QoS-aware data transfer

## Examples and Challenges (Cont'd)

- Middleware Services
  - E2E Data Replication Services
  - SDN-aware GridFTP
  - Combining Supercomputer Schedulers (like SLURM, PBS, LSF) with schedulers for SDN-aware data transfer
- Apps Services
  - Innovative and intelligent SDN-aware workflows
  - Interactive remote visualization
    - Application running on a supercomputer with coordinated remote visualization
  - Coordinated exascale computation and steering over multiple labs/sites
  - Tightly-coupled computation, data generation, storage, transmission and processing over multiple labs/sites
  - SDN-aware Fault-Tolerance and Job Migration