In the multi-core age, How do larger, faster and cheaper and more responsive memory sub-systems affect data management?

Panel at ADMS 2011

Dhabaleswar K. (DK) Panda

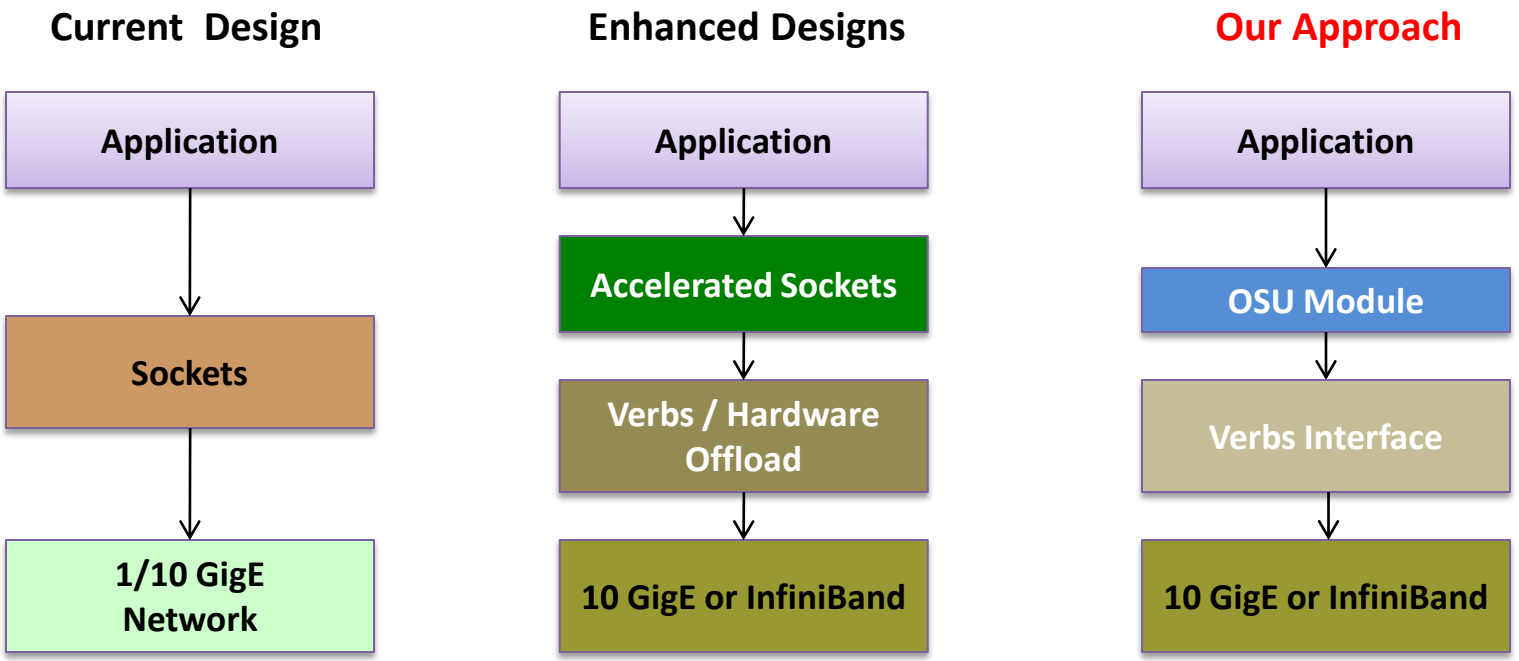
Network-Based Computing Laboratory
Department of Computer Science and Engineering
The Ohio State University

E-mail: panda@cse.ohio-state.edu

Motivation

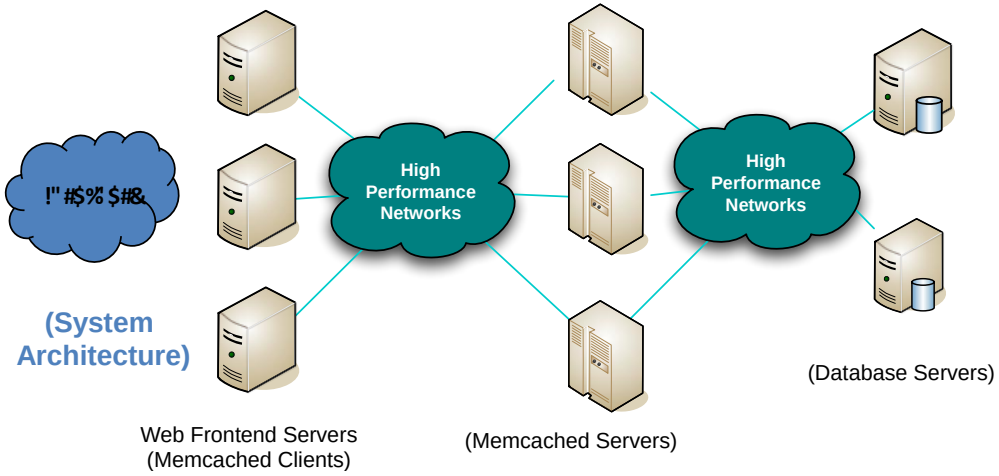
- Modern servers are providing us large amount of memory and multi-core processors per node-basis
- SSD is emerging as replacement for HDD
- Huge amount of memory across a set of servers provide new opportunities for designing data management systems?
- High performance commodity networks like InfiniBand with RDMA mechanism are allowing us to design very large HPC clusters with Petaflop performance
- Working on high-performance Message Passing Interface (MPI) software over InfiniBand (open-source MVAPICH project) for the last ten years
  - http://mvapich.cse.ohio-state.edu
  - Used by more than 1,650 organizations in 63 countries
  - Empowering many TOP500 systems and emerging Petaflop systems
    - 111,104-cores NASA Pleiades (7th ranked) and 62,976-core TACC Ranger (17th ranked)
  - Available with Redhat, SuSE and other Linux distros

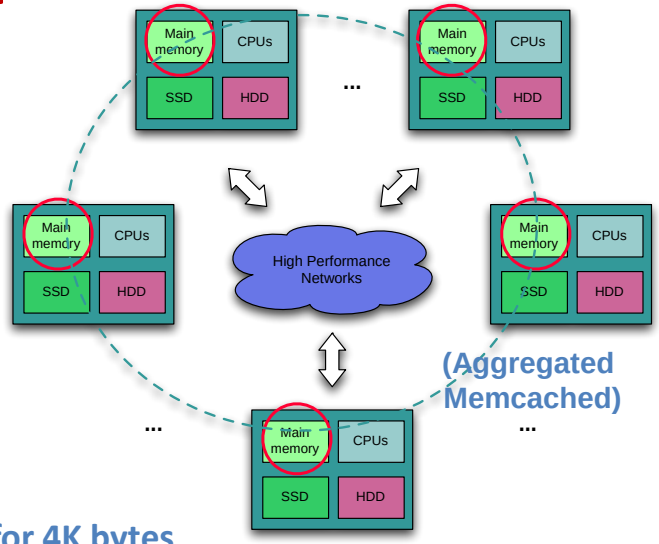
Can New Data Management Systems be designed with High-Performance Networks and Protocols?

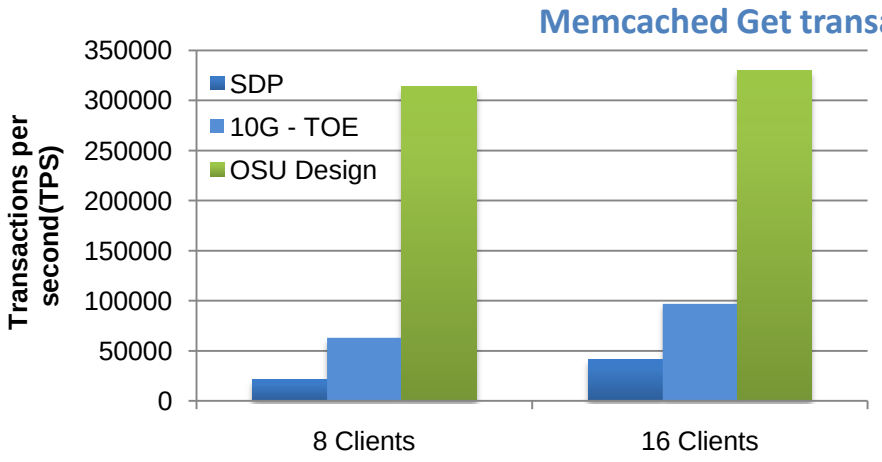


- Sockets not designed for high-performance
  - Stream semantics often mismatch for upper layers (Memcached, HBase, Hadoop)
  - Zero-copy not available for non-blocking sockets
- Interesting interplay between memory, storage and interconnect …

Memcached



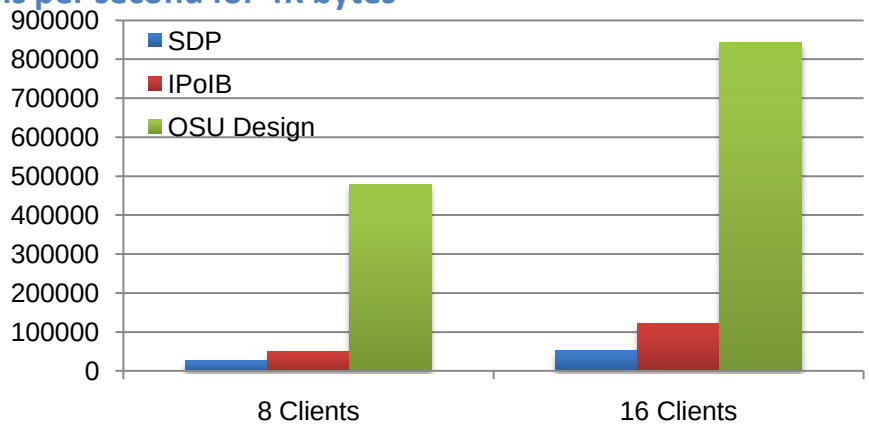




Intel Clovertown **Cluster (IB: DDR)**

On IB DDR about 330K/s for 16 clients

Almost factor of *four* improvement over 10GE (TOE)



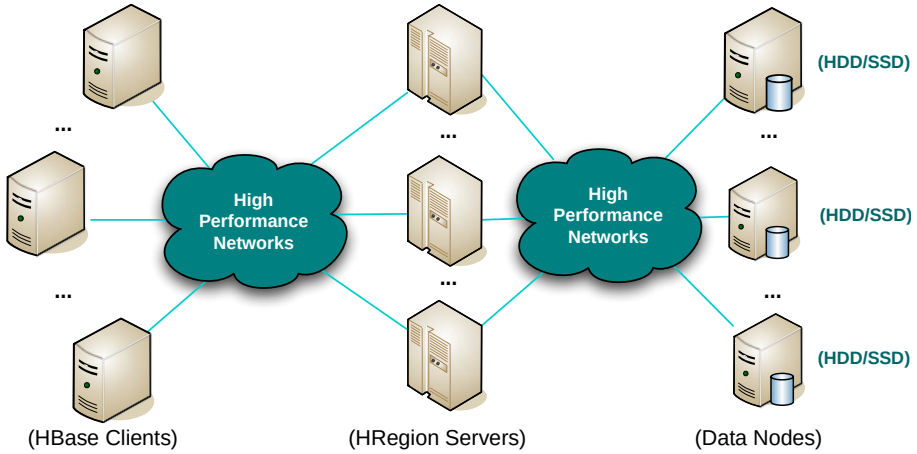
Intel Westmere Cluster (IB: QDR)

On IB QDR about 842K/s for 16 clients

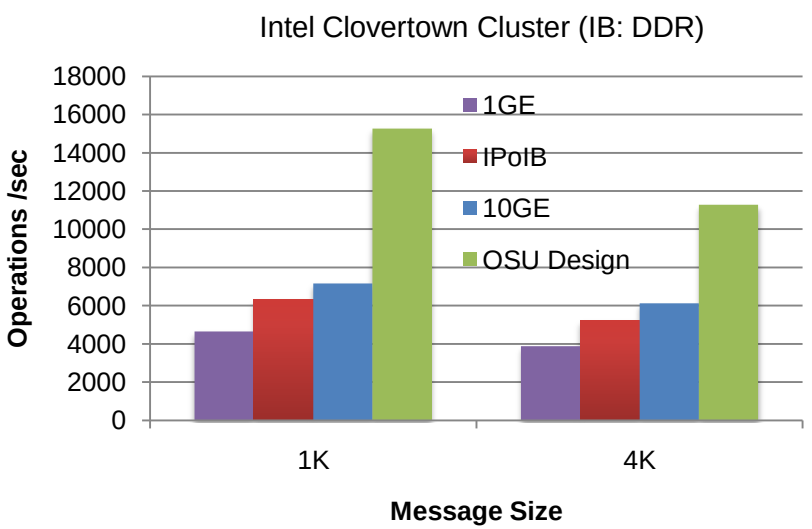
Almost factor of *seven* improvement over IPoIB

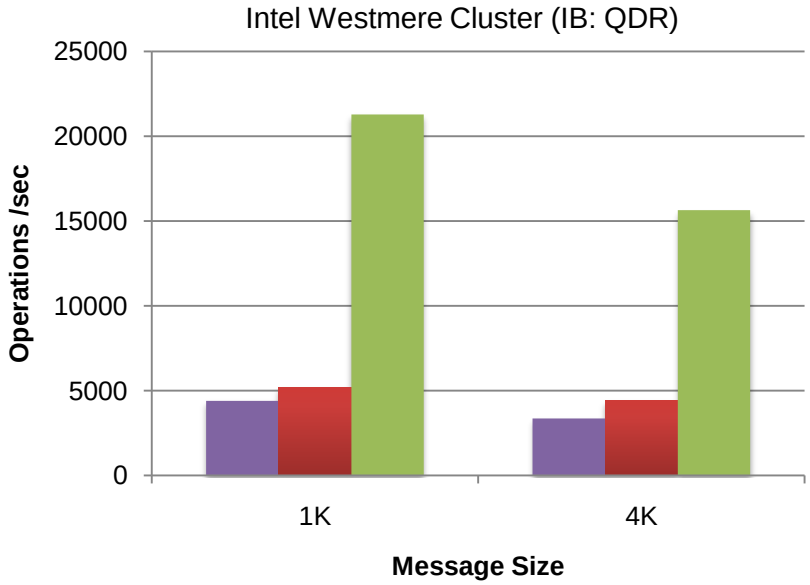
J. Jose, H. Subramoni, M. Luo, M. Zhang, J. Huang, M. W. Rahman, N. S. Islam, X. Ouyang, H. Wang, S. Sur and D. K. Panda, **Memcached Design on High Performance RDMA Capable Interconnects**, Int'l Conference on Parallel Processing (ICPP '11), Sept. 2011

HBase



(HBase System Architecture)



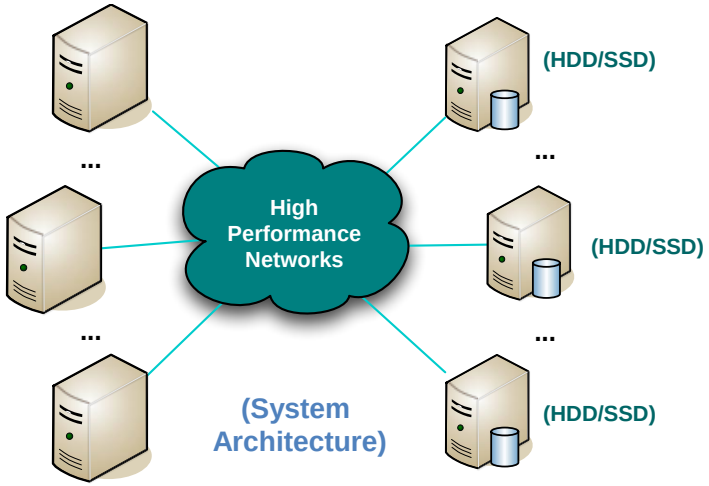


IB:DDR:

- **1K bytes – 65 us (15K TPS)**
- **4K bytes -- 88 us (11K TPS)**
- **Almost factor of *two* improvement over 10GE (TOE)**

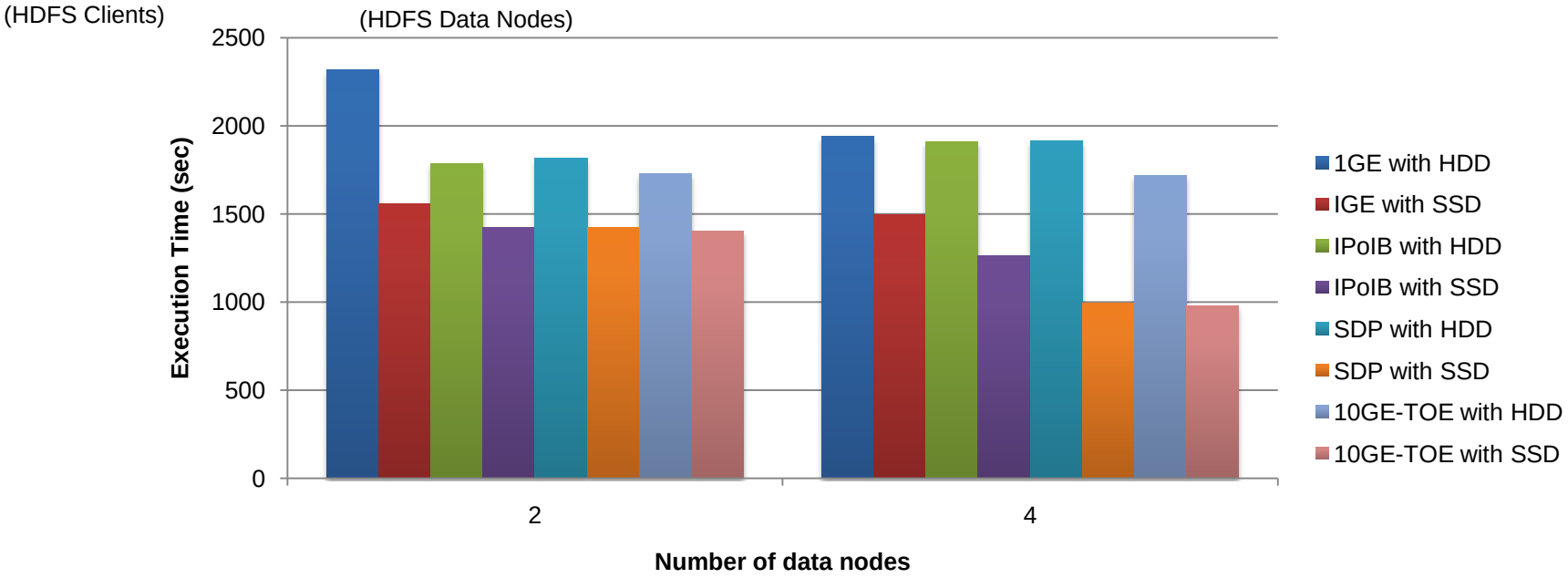
IB:QDR:

- **1K bytes – 47 us (22K TPS)**
- **4K bytes -- 64 us (16K TPS)**
- **Almost factor of *four* improvement over IPoIB for 1KB**



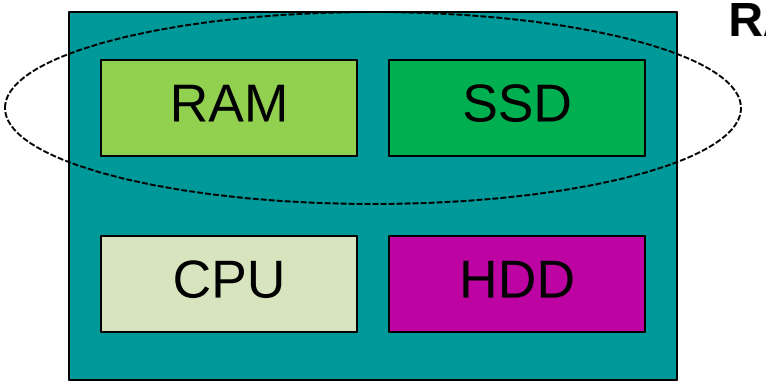
HDFS

- Sort: baseline benchmark for Hadoop
- Sort phase: I/O bound; Reduce phase: communication bound
- SSD improves performance by 28% using 1GigE with two DataNodes
- Benefit of 50% on four DataNodes using SDP, IPoIB or 10GigE



S. Sur, H. Wang, J. Huang, X. Ouyang and D. K. Panda, **Can High-Performance Interconnects Benefit Hadoop Distributed File System?, Workshop on Micro Architectural Support for Virtualization, Data Center Computing and Clouds**, in Conjunction with MICRO 2010, Dec 2010, Atlanta, GA, USA

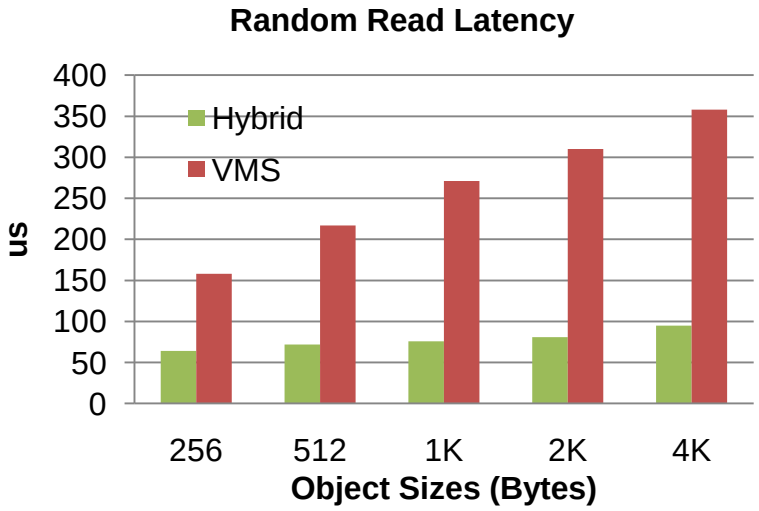
SSD-Assisted Hybrid Memory

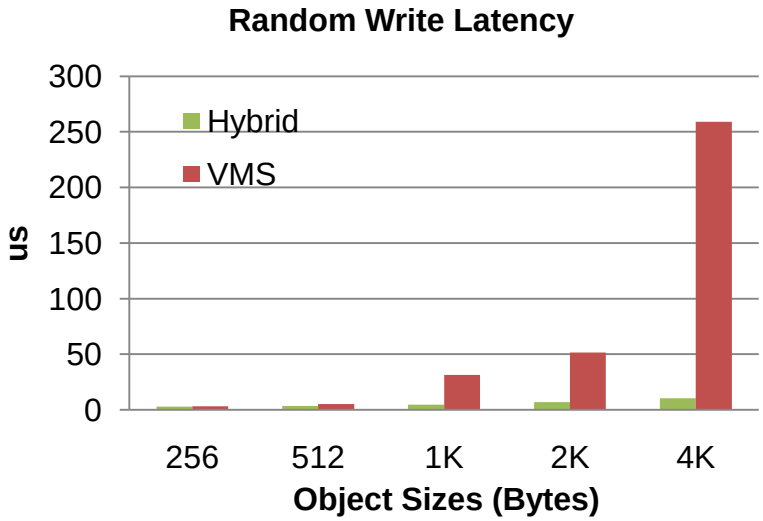


RAM/SSD Hybrid Memory

Random Read:
- **1 KB object:** Hybrid is 3.6X faster than VMS
- **4 KB object:** Hybrid is 3.8X faster than VMS

Random Write:
- **1 KB object:** Hybrid is 7.0X faster than VMS
- **4 KB object:** Hybrid is 24.7X faster than VMS

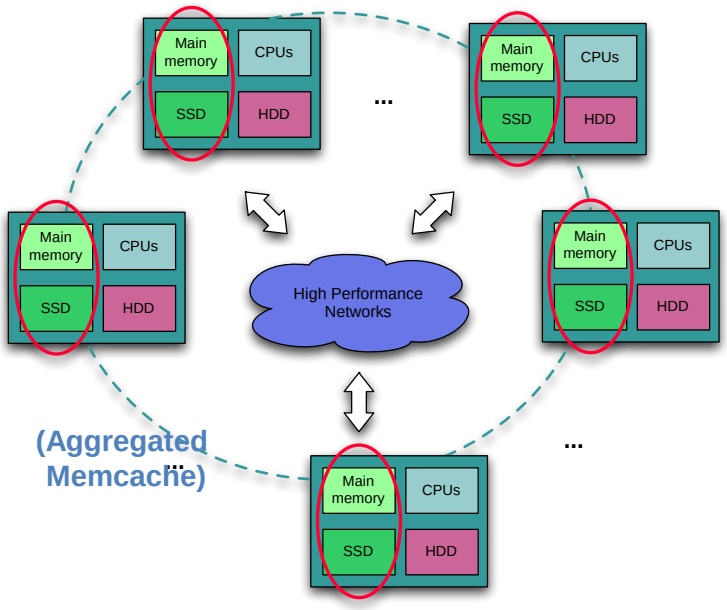


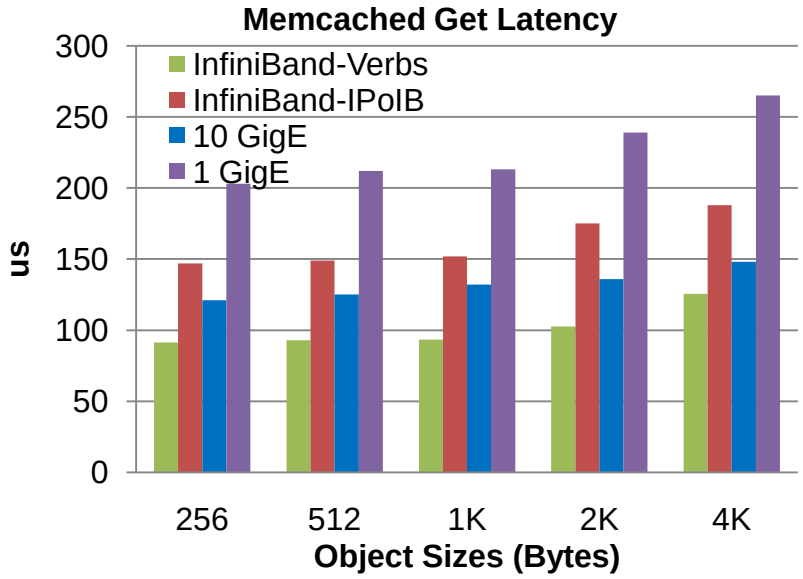


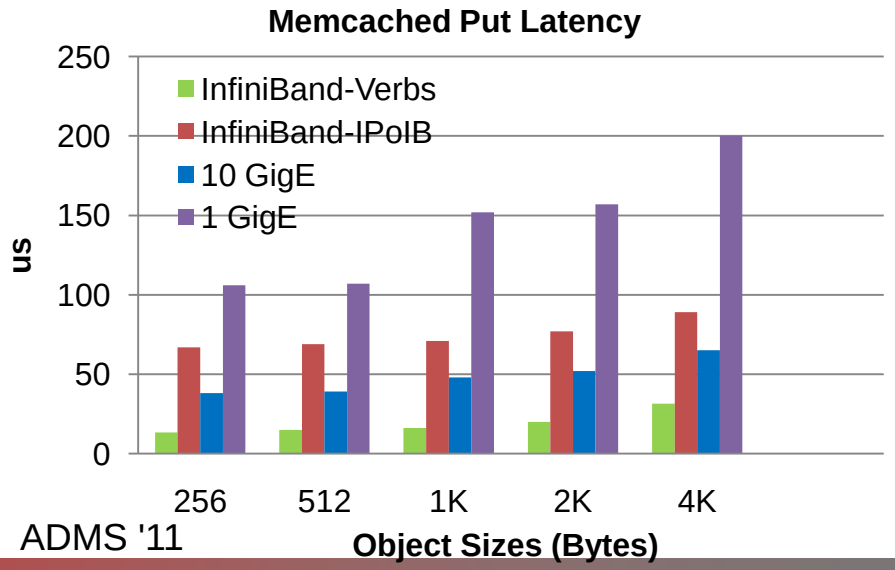
Hybrid: RAM/SSD Hybrid Memory
VMS: SSD as Virtual Memory Swap Device

SSD: Fusion-io ioDrive SLC 80GB

Memcached + Hybrid Memory







- **IB DDR, with Hybrid Memory**

Memcached Get with InfiniBand-Verbs:
- **1 KB object:** IB is 1.5X faster than 10GigE

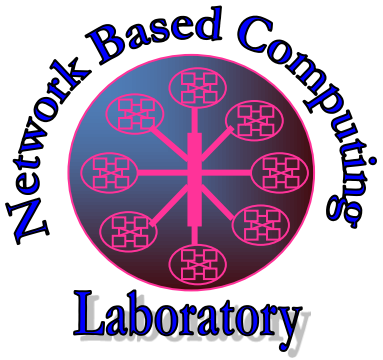
Memcached Put with InfiniBand-Verbs:
- **1 KB object:** IB is 2.9X faster than 10GigE

Conclusion

- High Performance networks like InfiniBand and RDMA protocols together with SSDs are opening up new ways to design modern enterprise systems
  - Aggregation of memory across nodes (Memcached)
  - Aggregation of memory and SSD (Hybrid memory with SSD in a node and Memcached + Hybrid memory)
  - High performance designs for HBase and HDFS
- Potential to design next-generation high-performance and scalable data management systems

Thank You!

panda@cse.ohio-state.edu



Network-Based Computing Laboratory

http://nowlab.cse.ohio-state.edu/

MVAPICH Web Page

http://mvapich.cse.ohio-state.edu/