Understanding and Optimizing Data Input/Output of Large-Scale Scientific Applications

Presented by

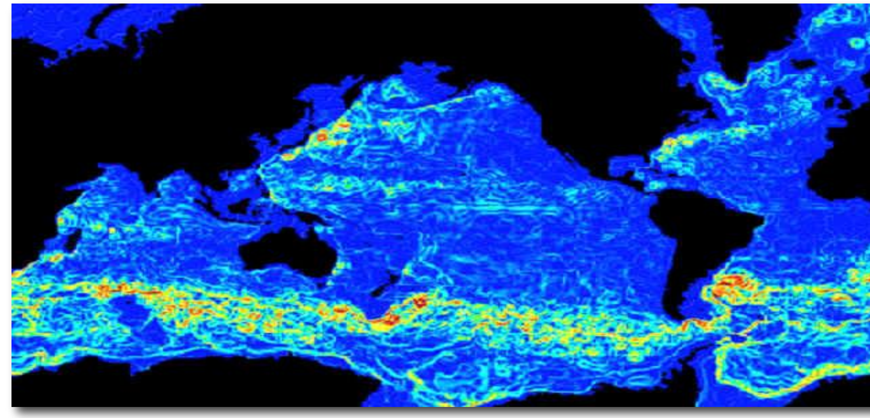
Jeffrey S. Vetter

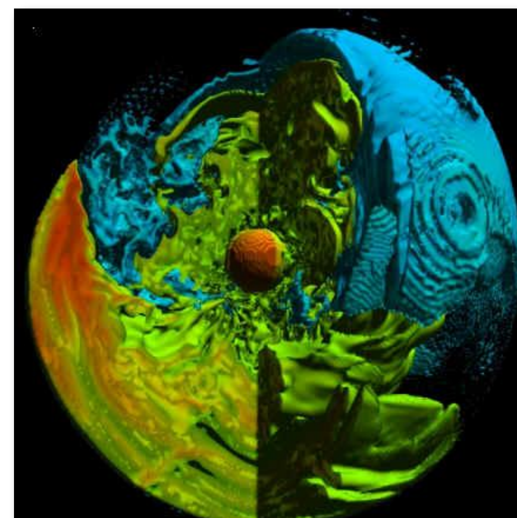
Leader
Future Technologies Group
Computer Science and Mathematics Division

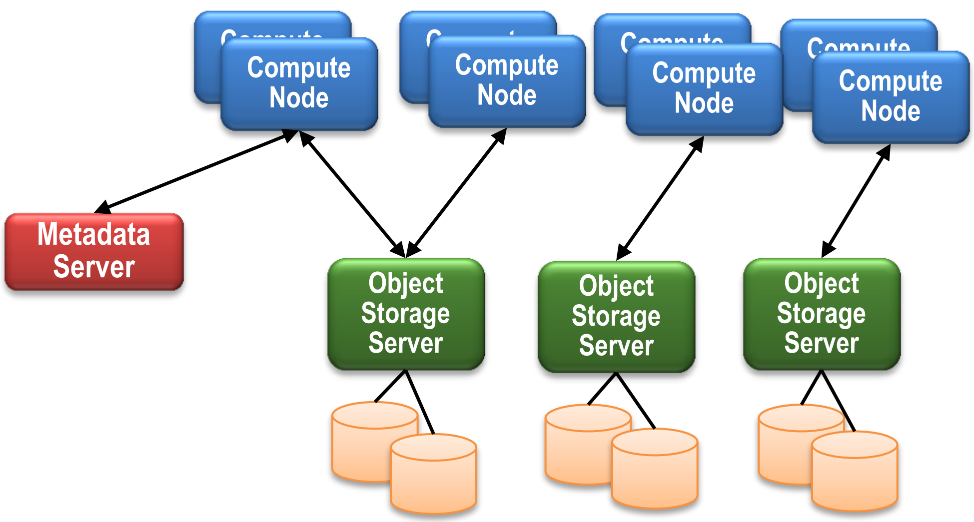
Team Members
Weikuan Yu, Yong Chen, Philip C. Roth

I/O for large-scale scientific computing

- Reading input and restart files
- Writing checkpoint files
- Writing movie, history files
- Gaps of understanding across domains; low efficiency

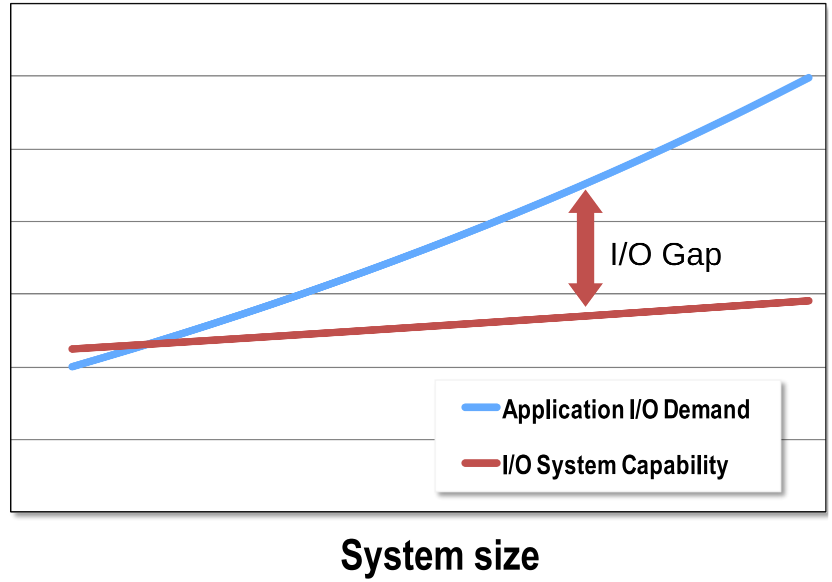




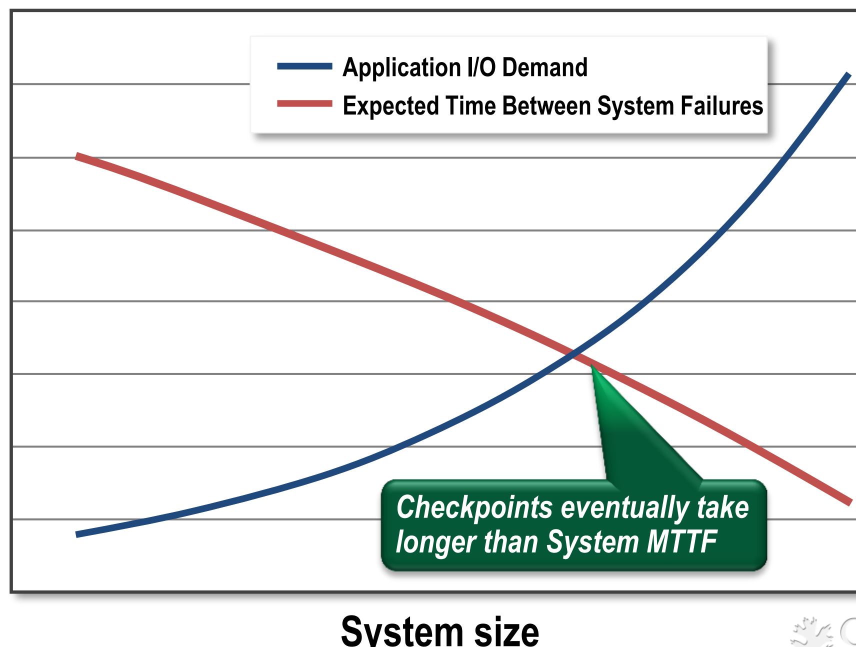


The I/O gap

- Widening gap between application I/O demands and system I/O capability

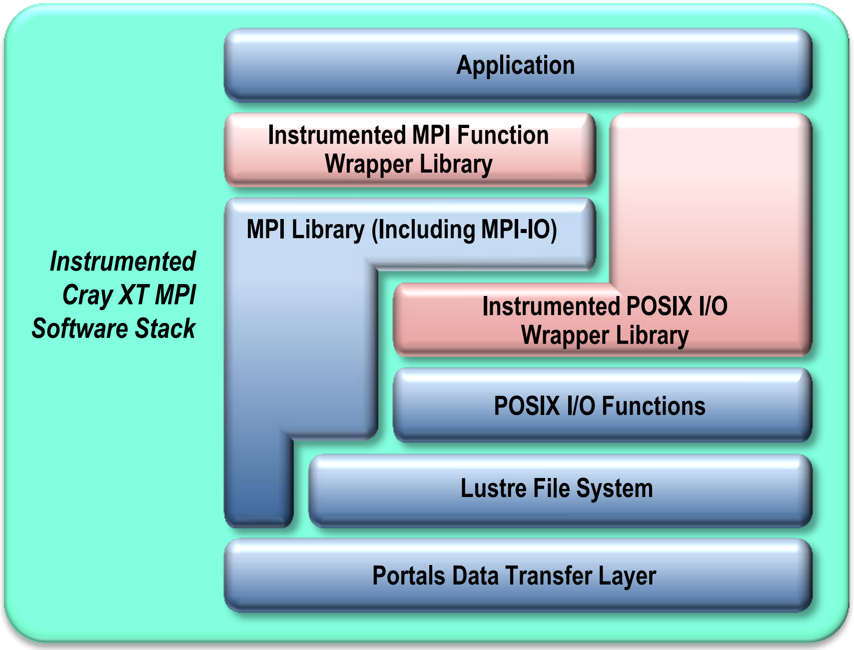


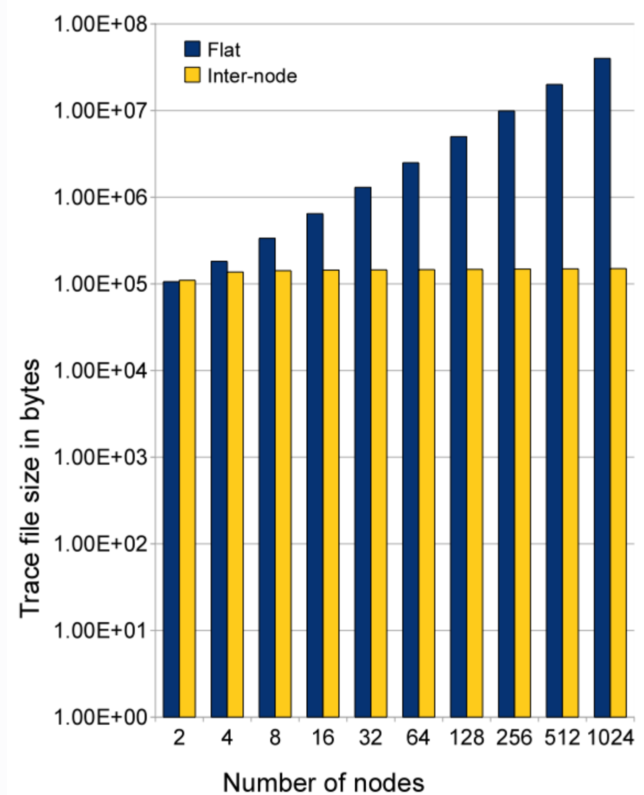
- Gap may grow too large for existing techniques (e.g., checkpointing) to be viable, due to decreases in system reliability as systems get larger



Insight into I/O behavior

- Scalable performance data collection infrastructure for Cray XT
- Gathers detailed I/O request data without changes to application source code
- Useful for
  - Characterizing application I/O
  - Driving storage system simulations
  - Deciding how and where to optimize I/O



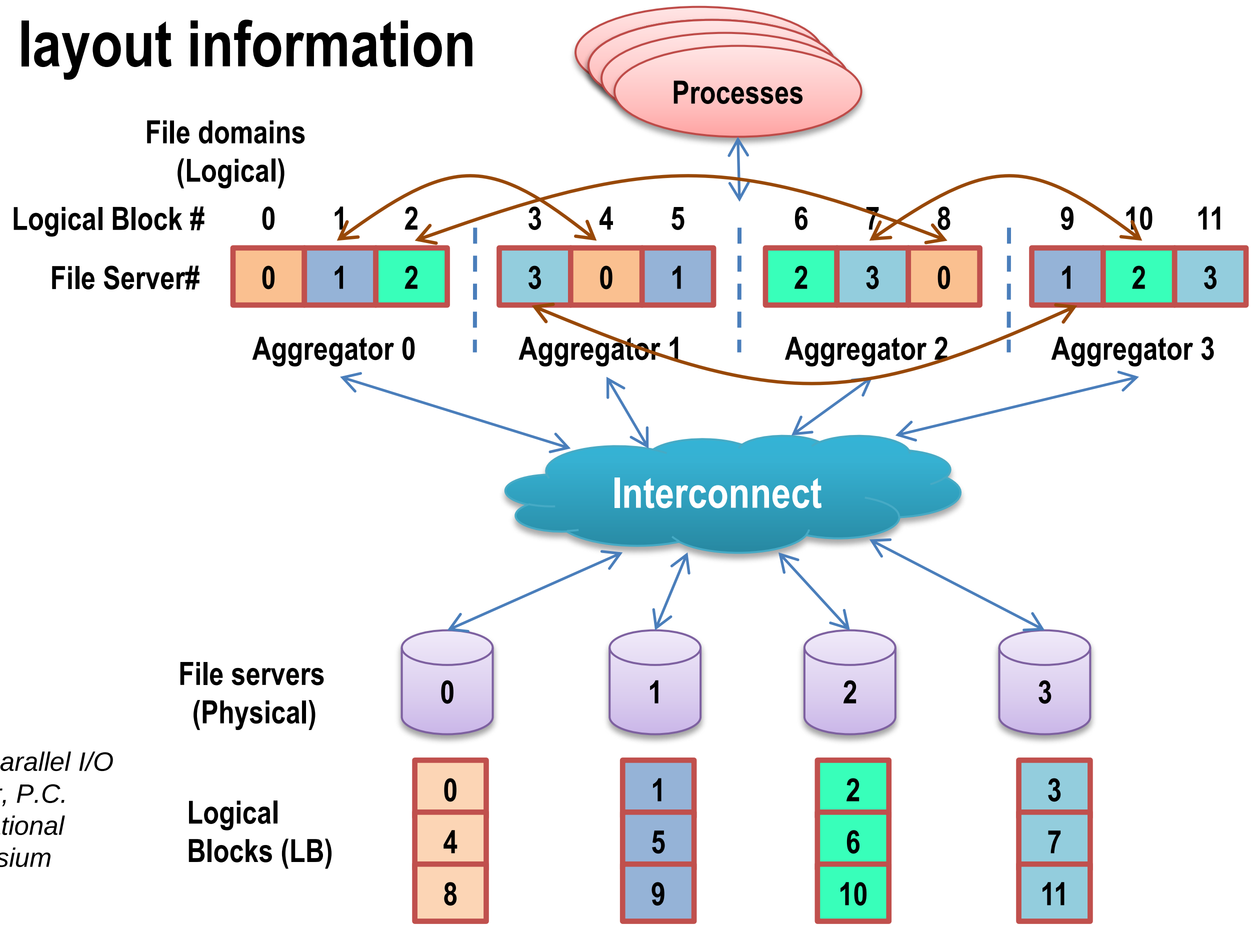


Event trace file size for FLASH I/O on Jaguar Cray XT4 at ORNL, with ScalaTrace compression (yellow) and without (blue). Note logarithmic Y axis.

Probabilistic Communication and I/O Tracing with Deterministic Replay at Scale, by X. Wu, K. Vijayakumar, F. Mueller, X. Ma, and P.C. Roth, in 2011 International Conference on Parallel Processing (ICPP 2011)

Layout-aware collective I/O

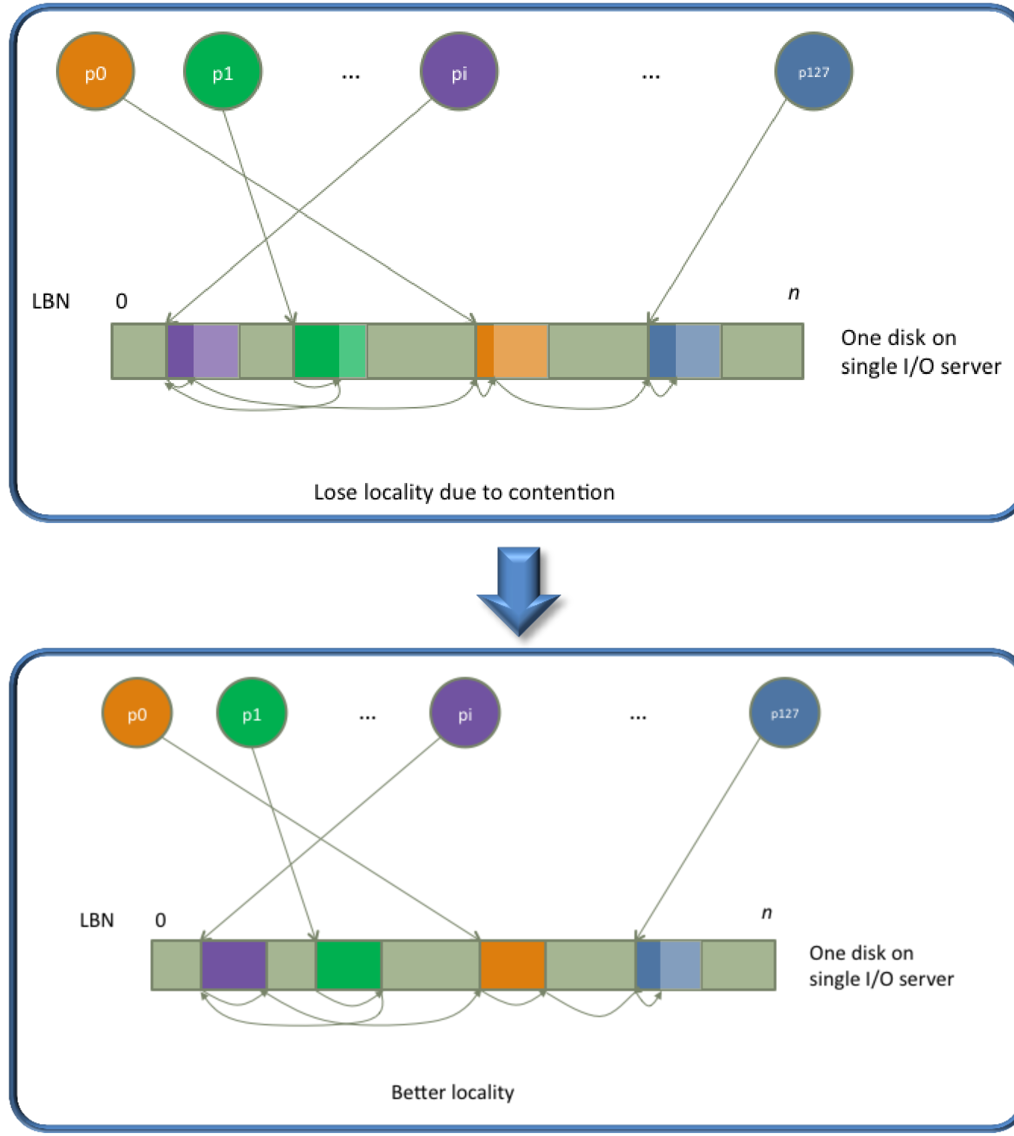
- Traditionally, parallel file systems and middleware are designed separately
- Exposing physical layout information to middleware allows it to reorder and reorganize accesses for better locality and improved performance



LACIO: A New Collective I/O Strategy for Parallel I/O Systems, by Y. Chen, X.-H. Sun, R. Thakur, P.C. Roth, and W.D. Gropp, In 25th IEEE International Parallel and Distributed Processing Symposium

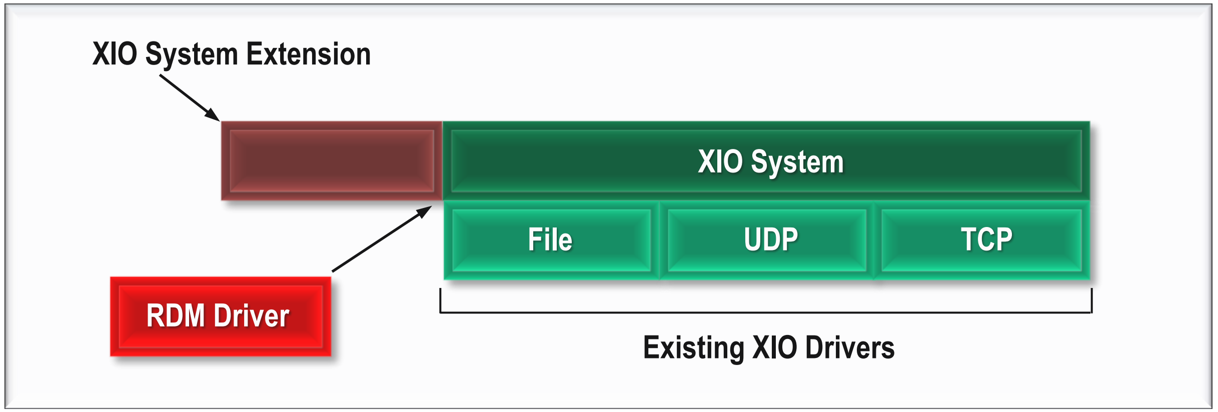
Layout-aware independent I/O

- Without awareness, independent accesses by multiple processes of a parallel application contend with each other
- With awareness, independent accesses serialized but do not contend with each other, giving better performance to application as a whole

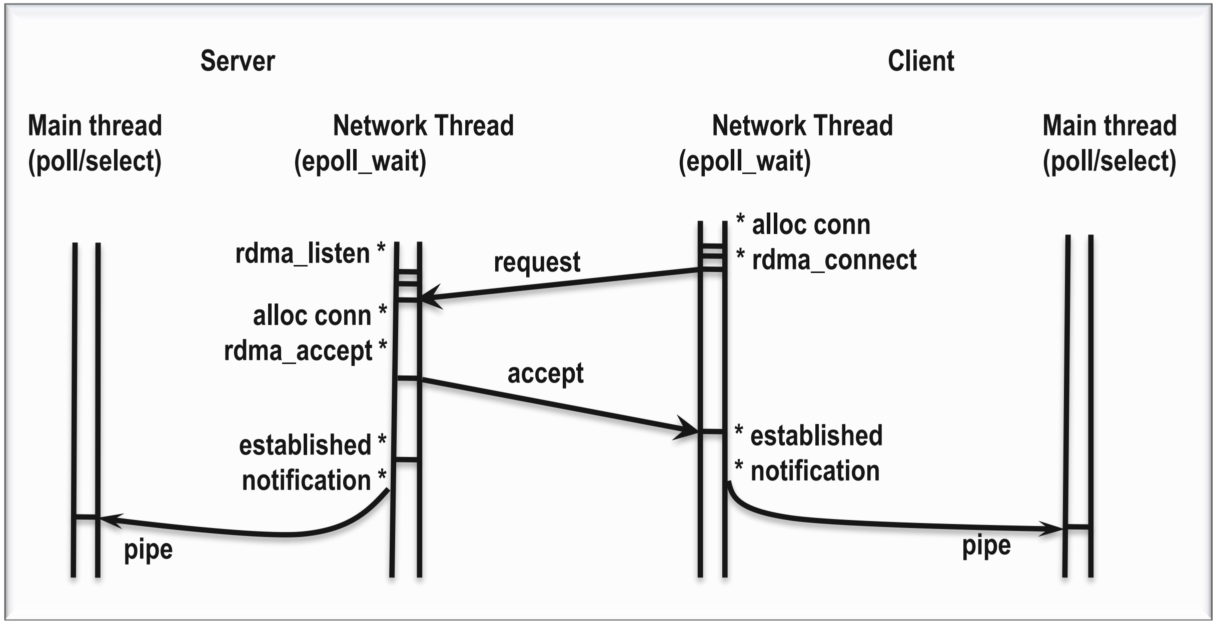


RXIO: High performance GridFTP on InfiniBand

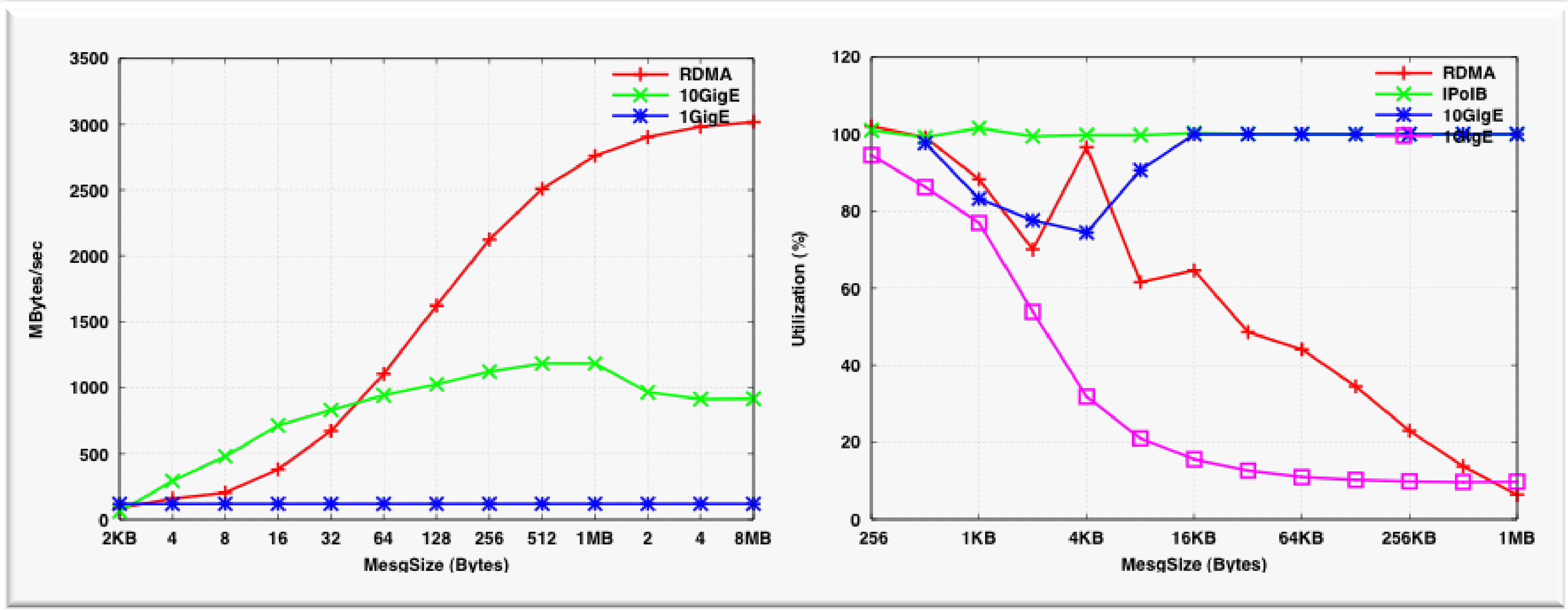
Software Architecture



Connection Establishment



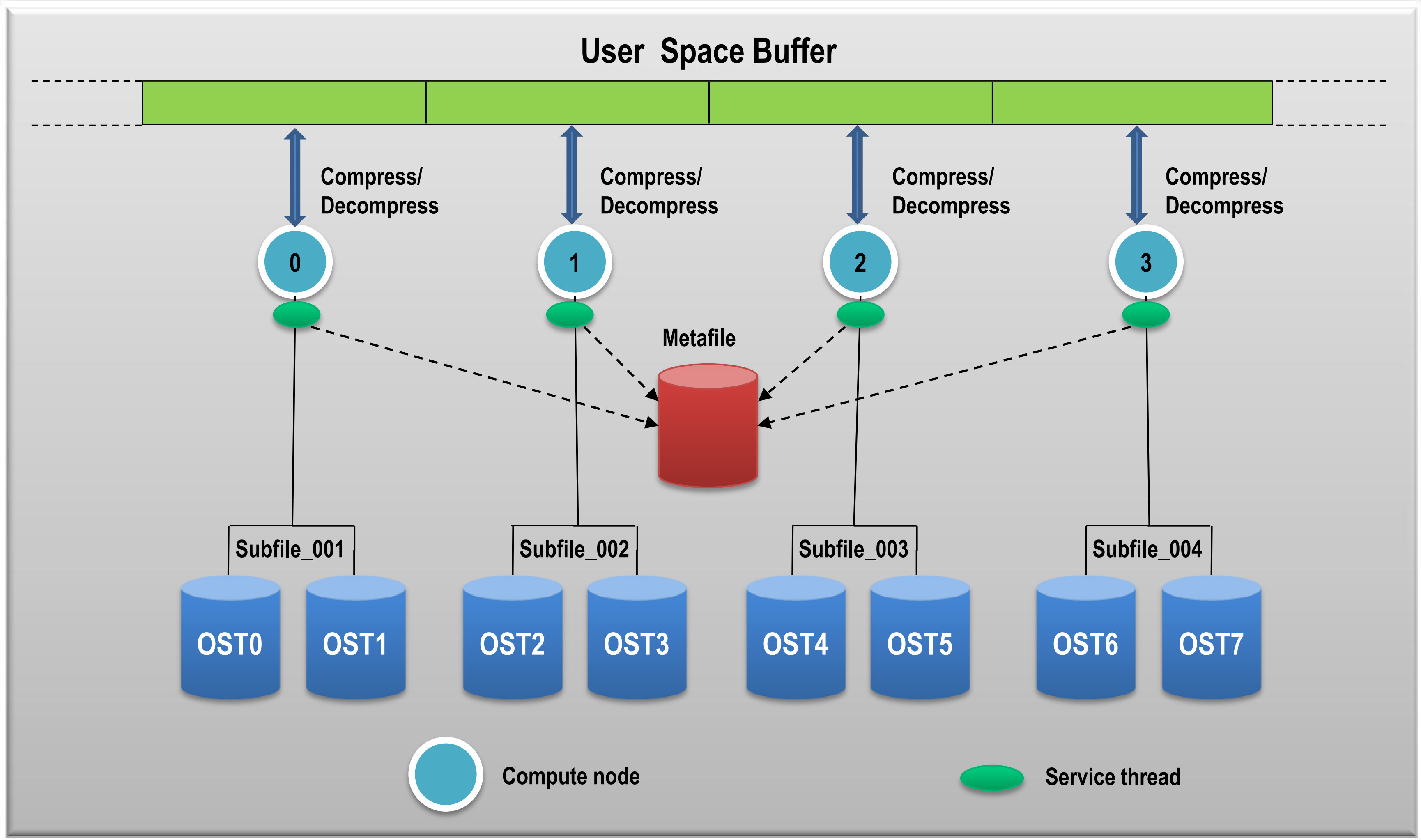
Performance benefits of RXIO



- Improve GridFTP bandwidth by three times compared to 10GigE
- At the same time, dramatically reduce the CPU utilization

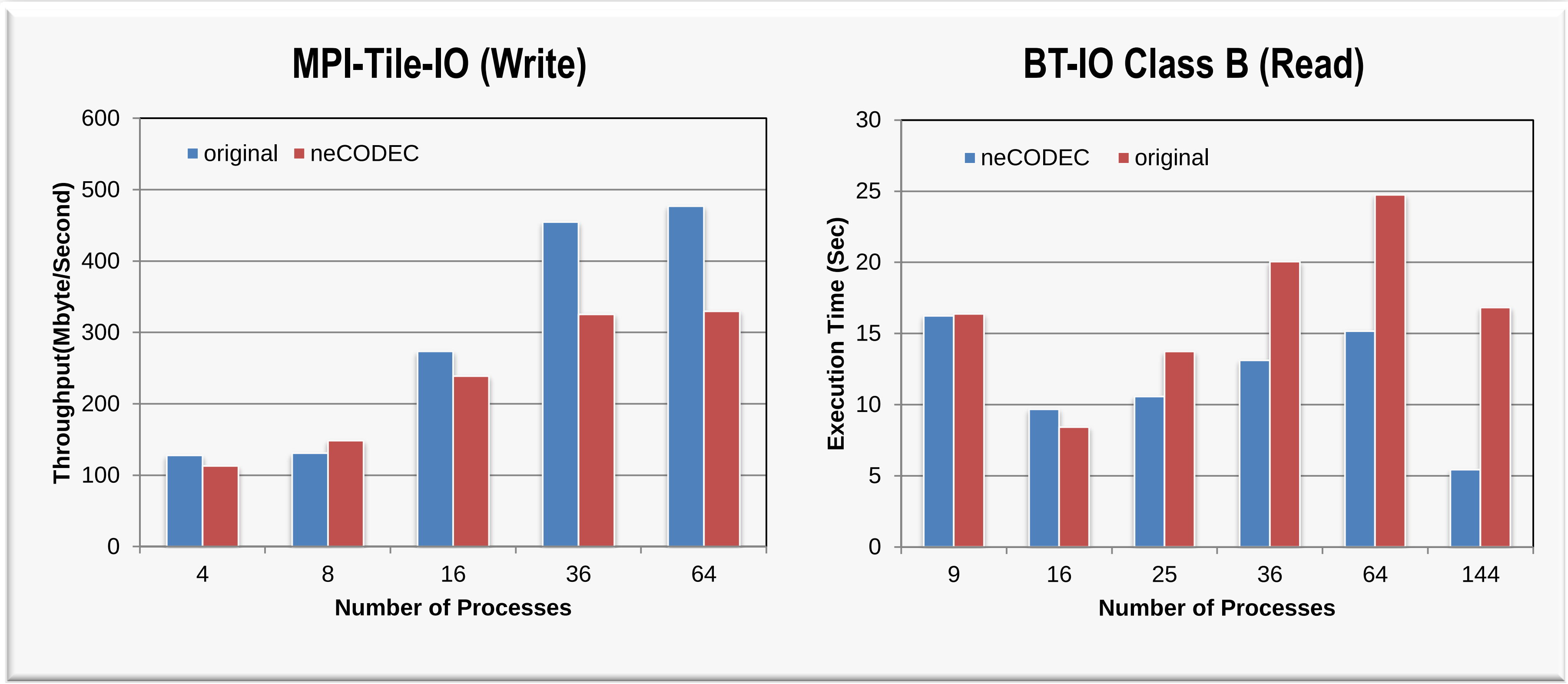
Efficient Zero-Copy Noncontiguous I/O for Globus on InfiniBand by W. Yu, Y. Tian, J.S. Vetter. In Proceedings of the Third International Workshop on Parallel Programming Models and Systems Software for High-End Computing (P2S2¹10), San Diego, CA.

neCODEC: Nearline data compression for scientific applications



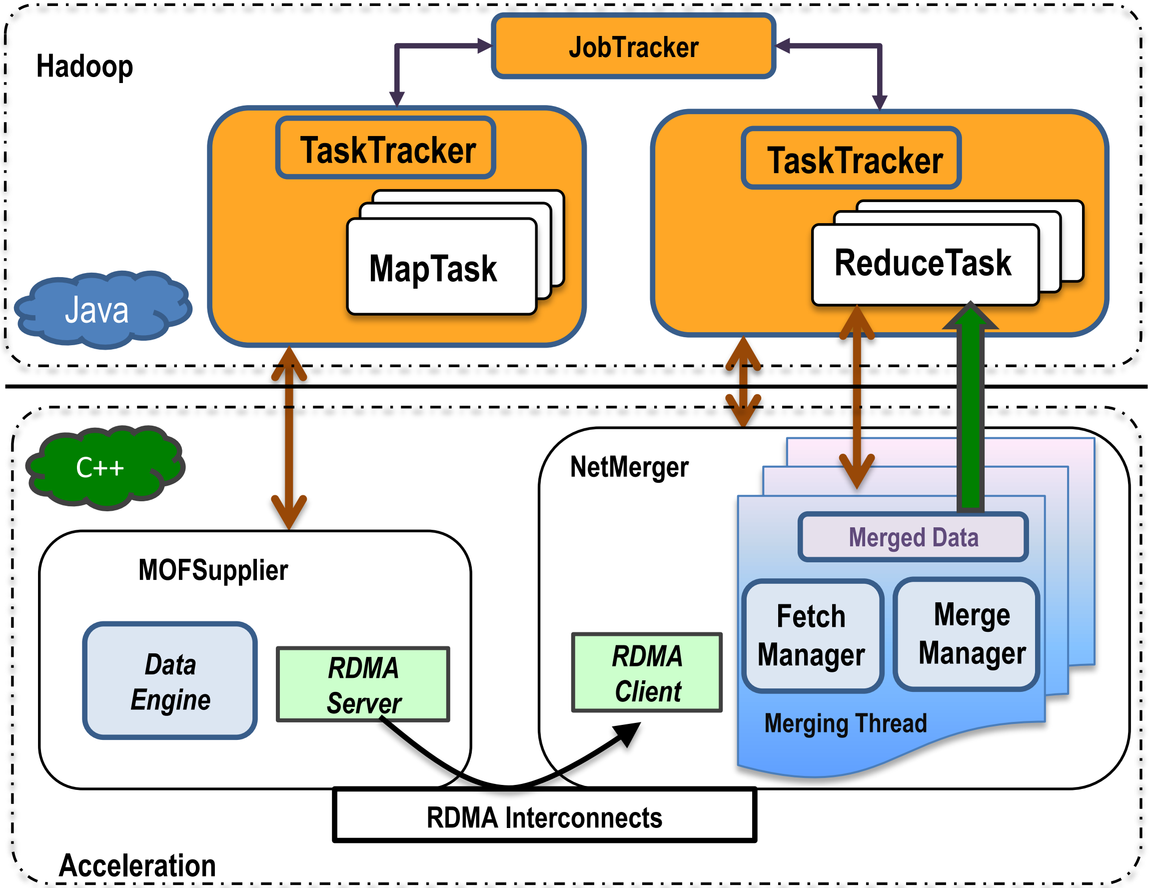
neCODEC: Nearline Data Compression for Data-Intensive Parallel Applications, by Y. Tian, W. Yu, J.S. Vetter, H. Liu. In review.

Performance results of neCODEC



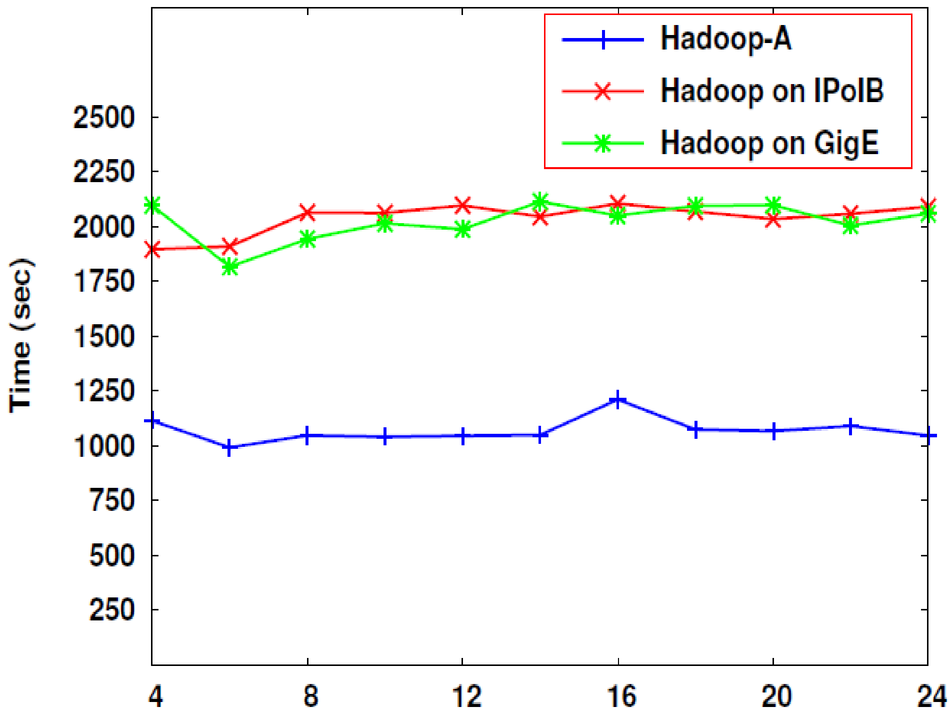
neCODEC improves the read and write bandwidth for MPI-Tile-IO and BT-IO

Hadoop Acceleration – UDA (Unstructured Data Accelerator)

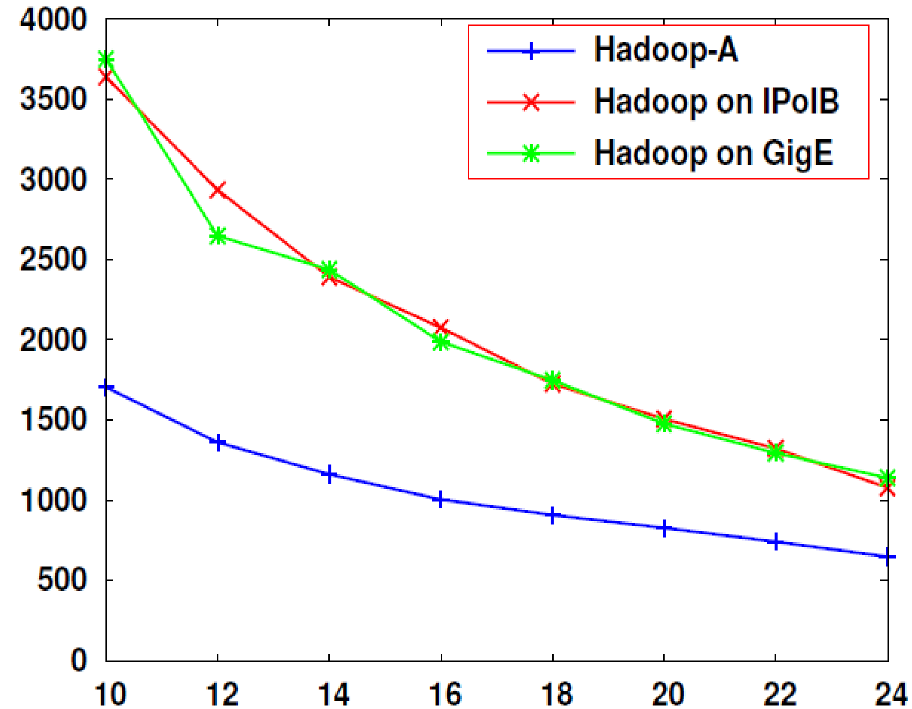


Yandong Wang, Xinyu Que, Weikuan Yu, Dror Goldenberg, Dhiraj Sehgal. *Hadoop Acceleration through Network Levitated Merging. SC11. Seattle, WA.*

Data Processing Scalability with UDA

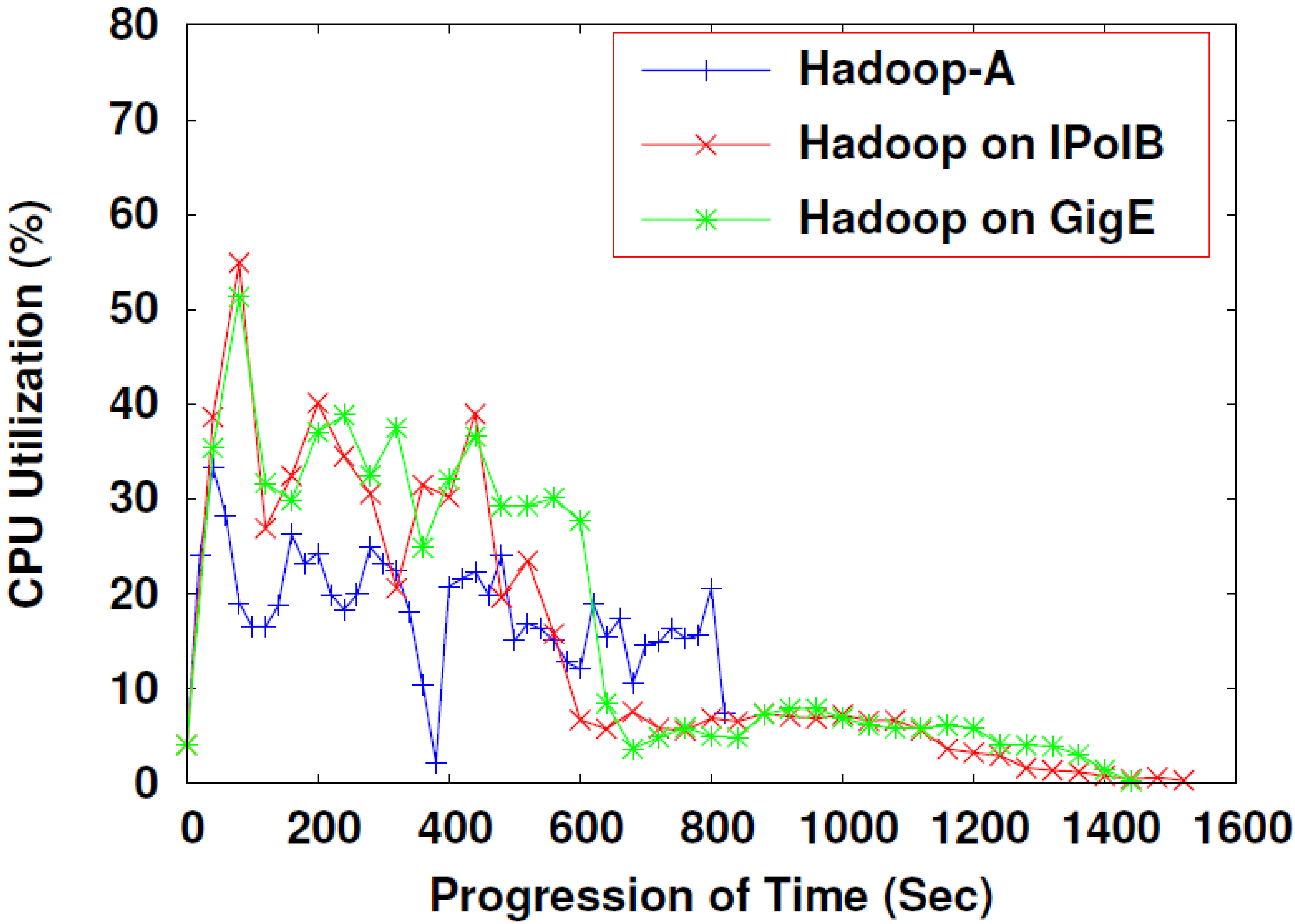


Execution Time with Fixed Dataset Per Reducer



Execution Time with Fixed Data Size Per Job

Reduced CPU Utilization with UDA



Contacts

Jeffery S. Vetter

Leader
Future Technologies Group
Computer Science and Mathematics Division
(865) 356-1649
vetter@ornl.gov

Weikuan Yu

ORNL/Auburn University
(344) 844-6330
wkyu@auburn.edu

Yong Chen

Texas Tech University(Formerly ORNL Postdoc)
(806) 742-3527 x230
yong.chen@ttu.edu

Philip C. Roth

(865) 241-1543
rothpc@ornl.gov

For more information see
http://ft.ornl.gov/projects/io/

U.S. DEPARTMENT OF ENERGY

SciDAC Scientific Discovery through Advanced Computing

U.S. DEPARTMENT OF ENERGY Office of Science

OAK RIDGE National Laboratory