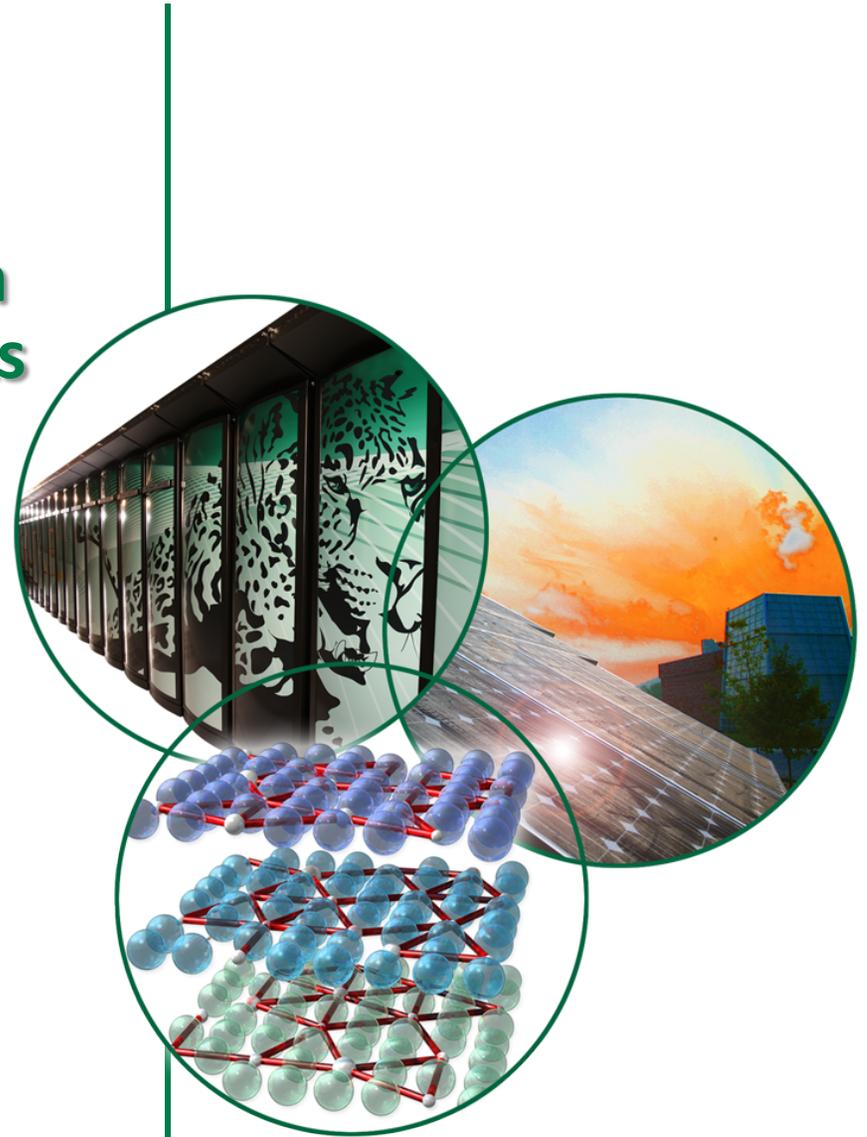


Toward Performance Prediction of Tree-Based Overlay Networks on the Cray XT

Philip C. Roth

Computer Science and Mathematics Division
Oak Ridge National Laboratory



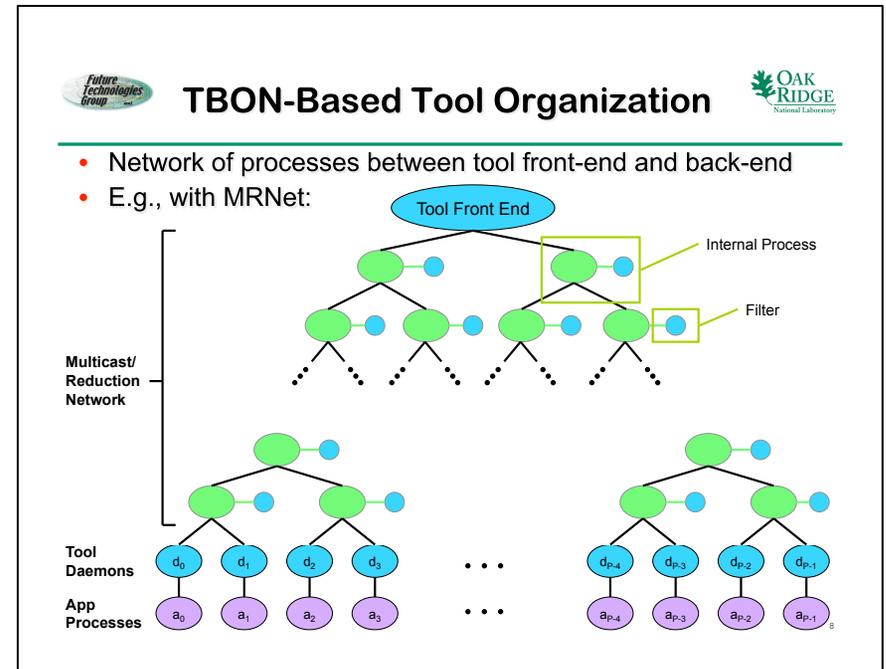
U.S. DEPARTMENT OF
ENERGY



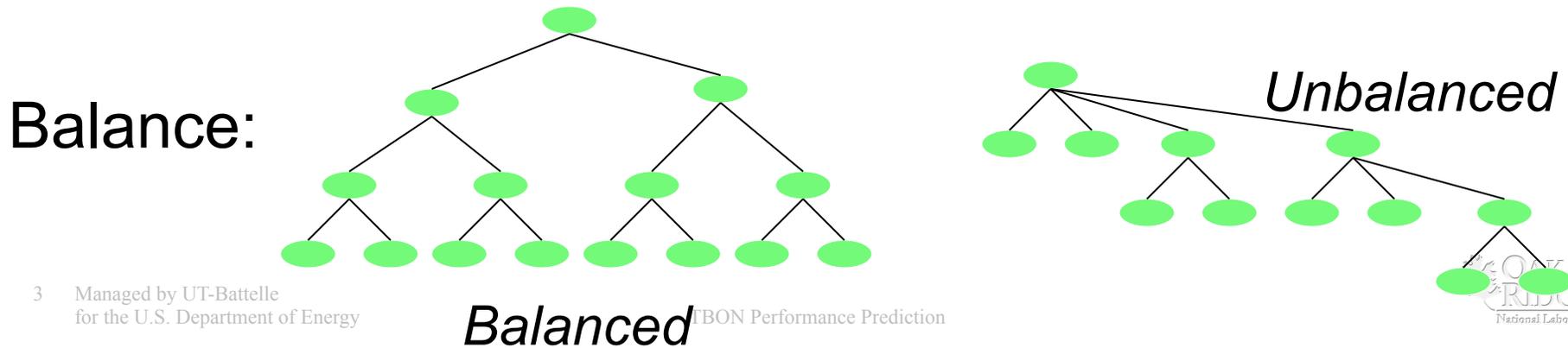
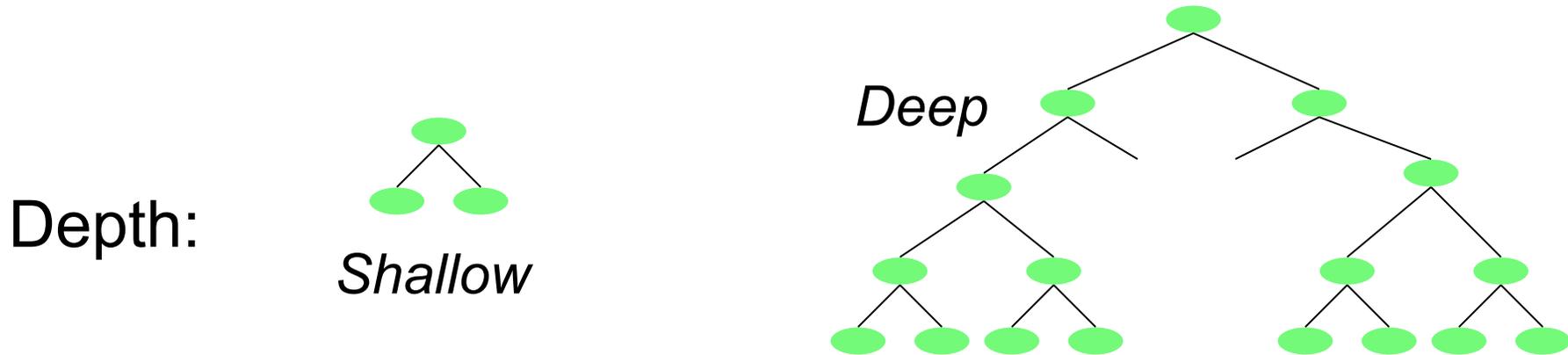
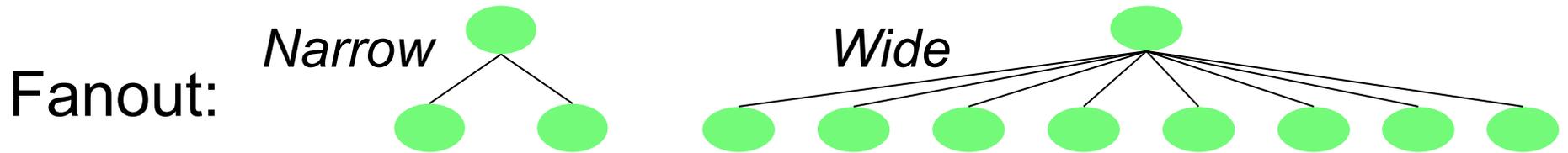
 **OAK RIDGE NATIONAL LABORATORY**
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

TBONs and MRNet

- A Tree-Based Overlay Network (TBON) like MRNet provides scalable infrastructure for tools and applications
- MRNet's process topology and placement support is extremely flexible (on most platforms)
 - Any tree topology
 - Internal processes on same nodes as application processes, or on distinct nodes



TBON Topology Flexibility



The Problem With Flexibility

- Flexibility leads to questions identifying “best” process topology and placement
- Interaction of several factors determine “best”
 - Performance (tool and application)
 - System hardware and software
 - Purpose
 - Even economics (e.g., can I afford to request “extra” nodes for MRNet processes given my allocation budget?)
- Decision process often not rigorous – using “rule of thumb”

TBON Performance Prediction

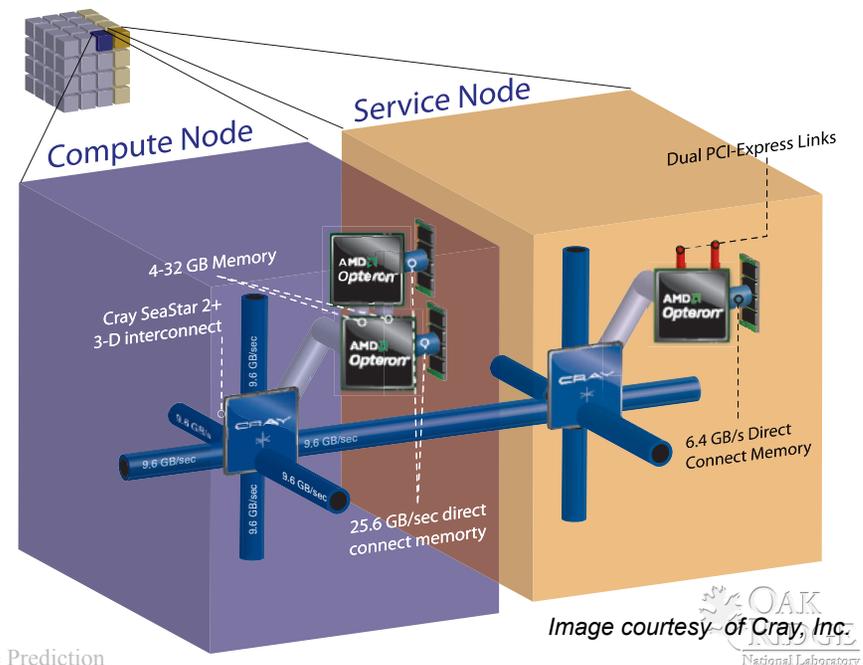
- Goal: Given a node allocation on a leadership class system, to be able to identify “best” MRNet process placement and topology
- Several constraints:
 - Tool multicast and reduction requirements
 - Behavior of application under study
 - Other activity on the system
 - System software and hardware

Target Platform

- Cray XT is target platform
 - Jaguar XT5 and XT4 systems at Oak Ridge National Laboratory (ORNL)
 - Hopper XT5 at NERSC
 - Kraken XT5 at ORNL
- Opteron-based nodes arranged in 3D mesh with possibility of torus links
- Cray Linux Environment



Managed by UT-Battelle
for the U.S. Department of Energy
Image courtesy of the National Center of Computational Sciences,
Oak Ridge National Laboratory



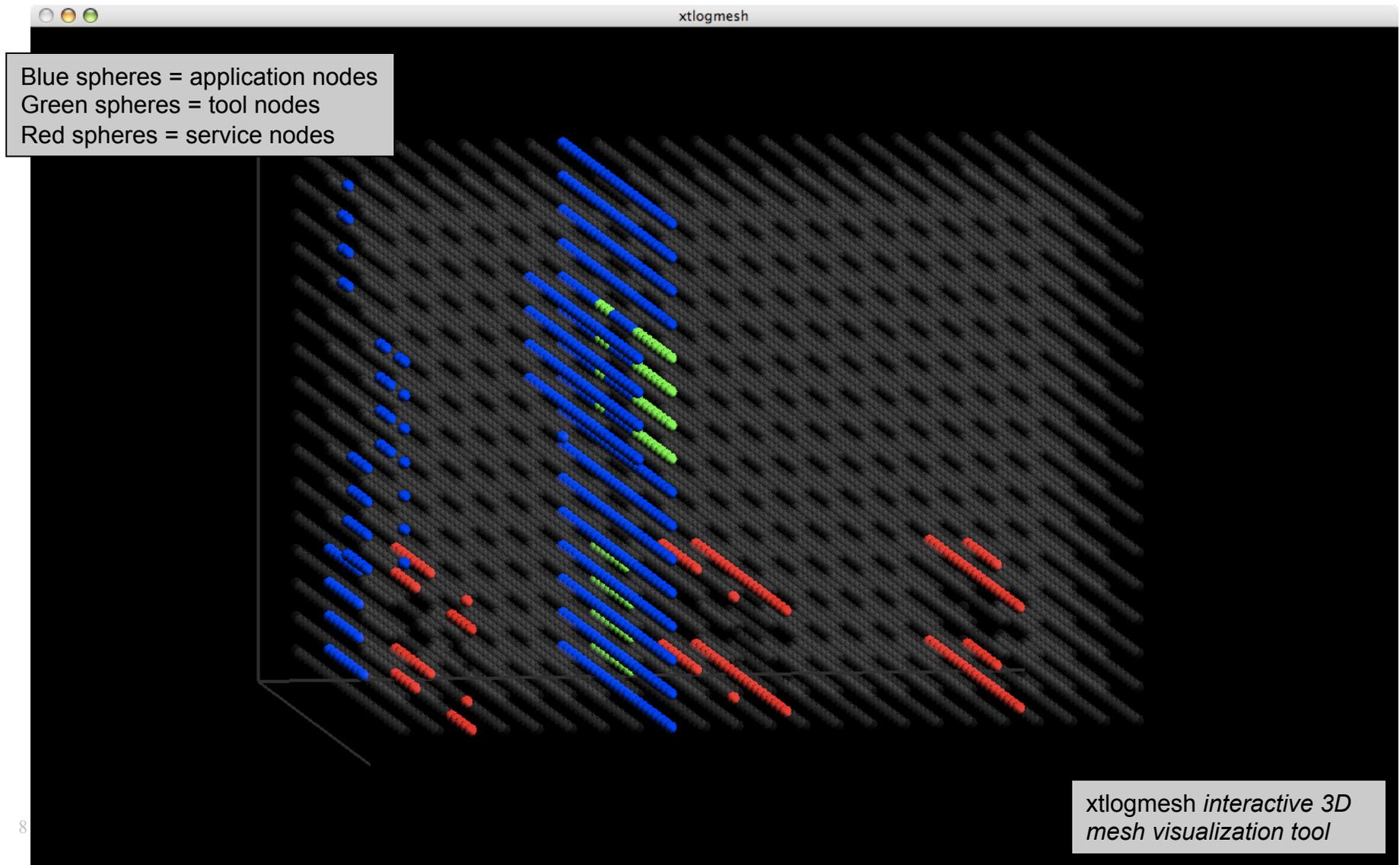
TBON Performance Prediction

Image courtesy of Cray, Inc.
OAK
AND
NERSC
National Laboratory

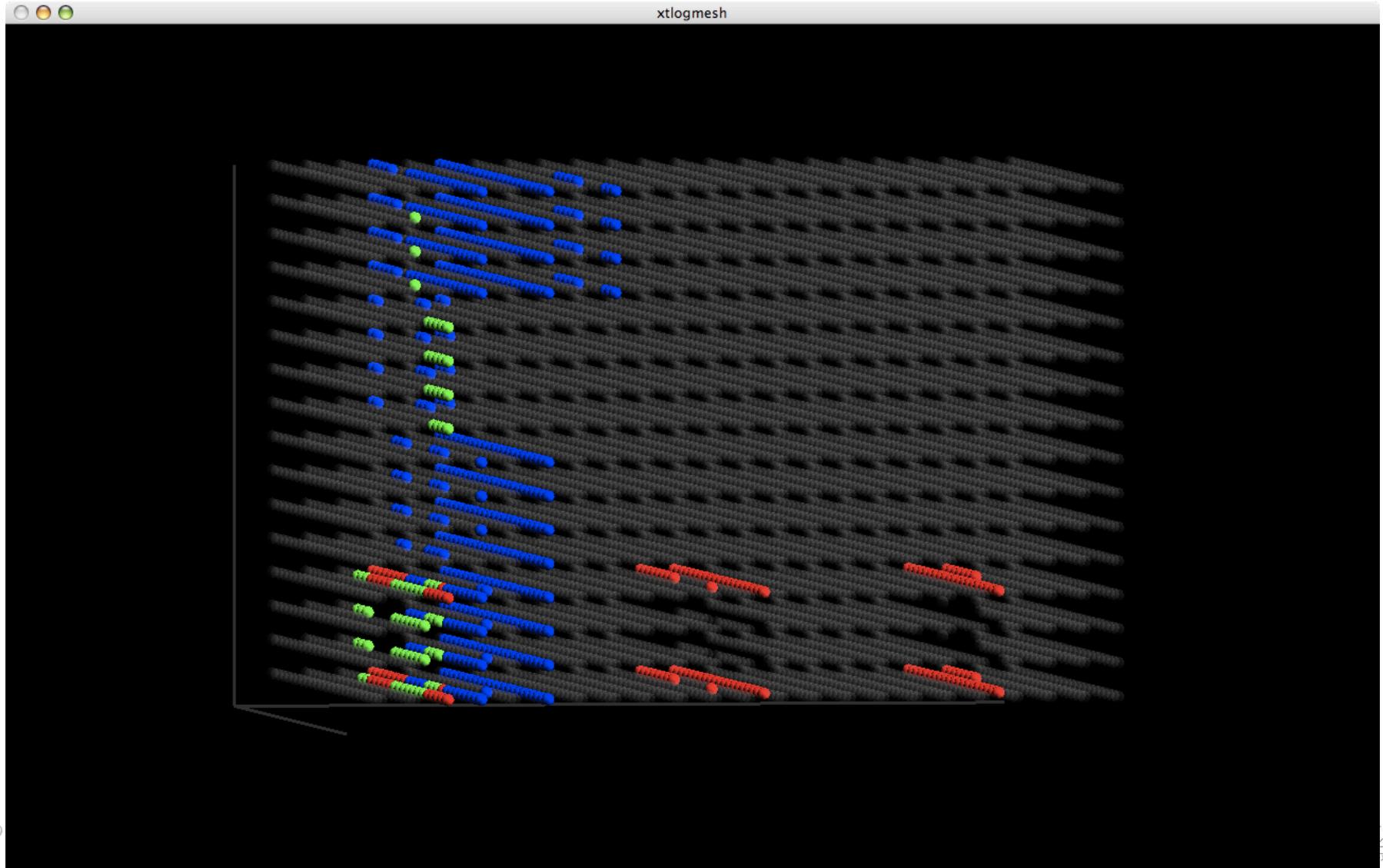
Cray XT TBON Process Placement Tests

- Goal: understand Cray XT allocation characteristics & their impact on MRNet-based tool process placement
- Used simple MPI/Portals program to collect node number and position within the XT mesh
 - Done on earlier generation ORNL Jaguar with dual-core Opterons
- Batch job launched two *independent* instances of the program:
 - 512 application nodes (1024 application processes)
 - 72 tool nodes (enough for balanced 8-way TBON topology, assuming front-end is on batch script or login node)

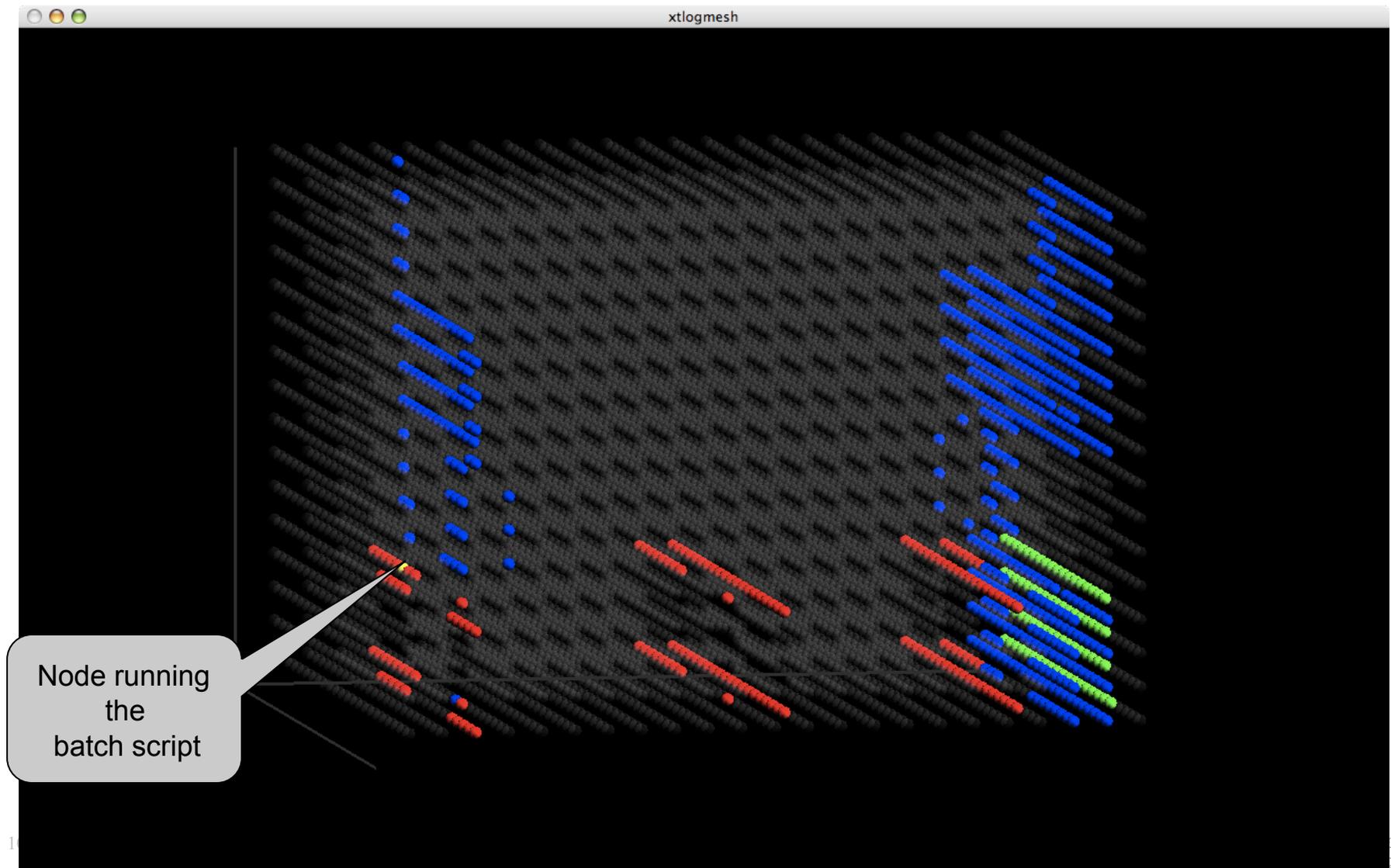
Jaguar Placement Trial 1



Jaguar Placement Trial 2

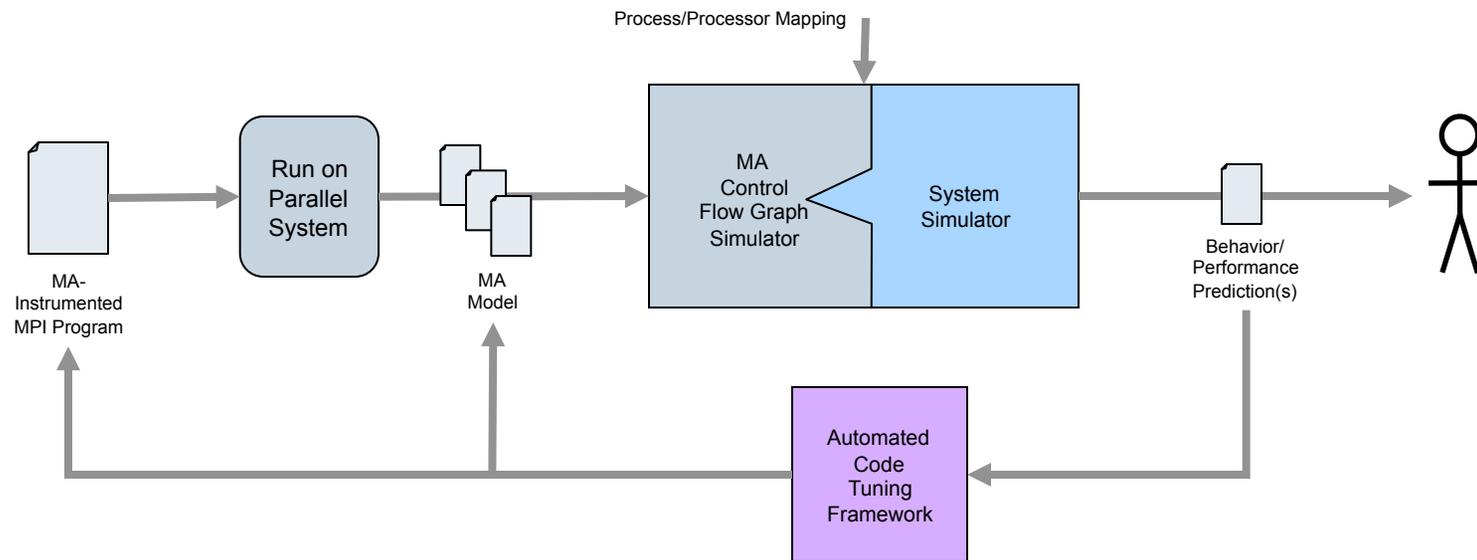


Jaguar Placement Trial 3



Our Approach

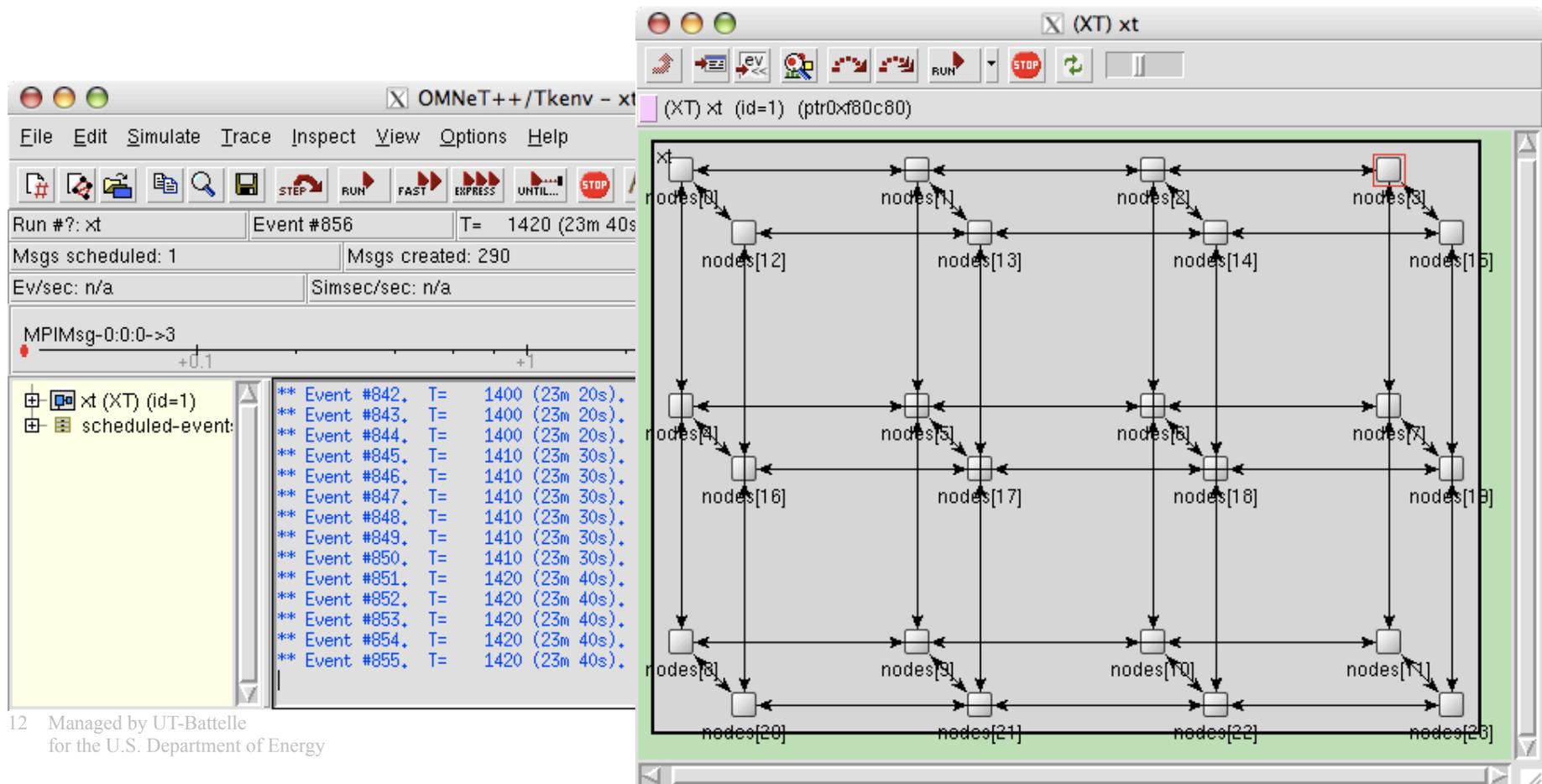
- Discrete event simulation of XT system nodes running app and MRNet processes
- Builds on work on Modeling Assertions, Simulation, and Tuning (MAST) framework



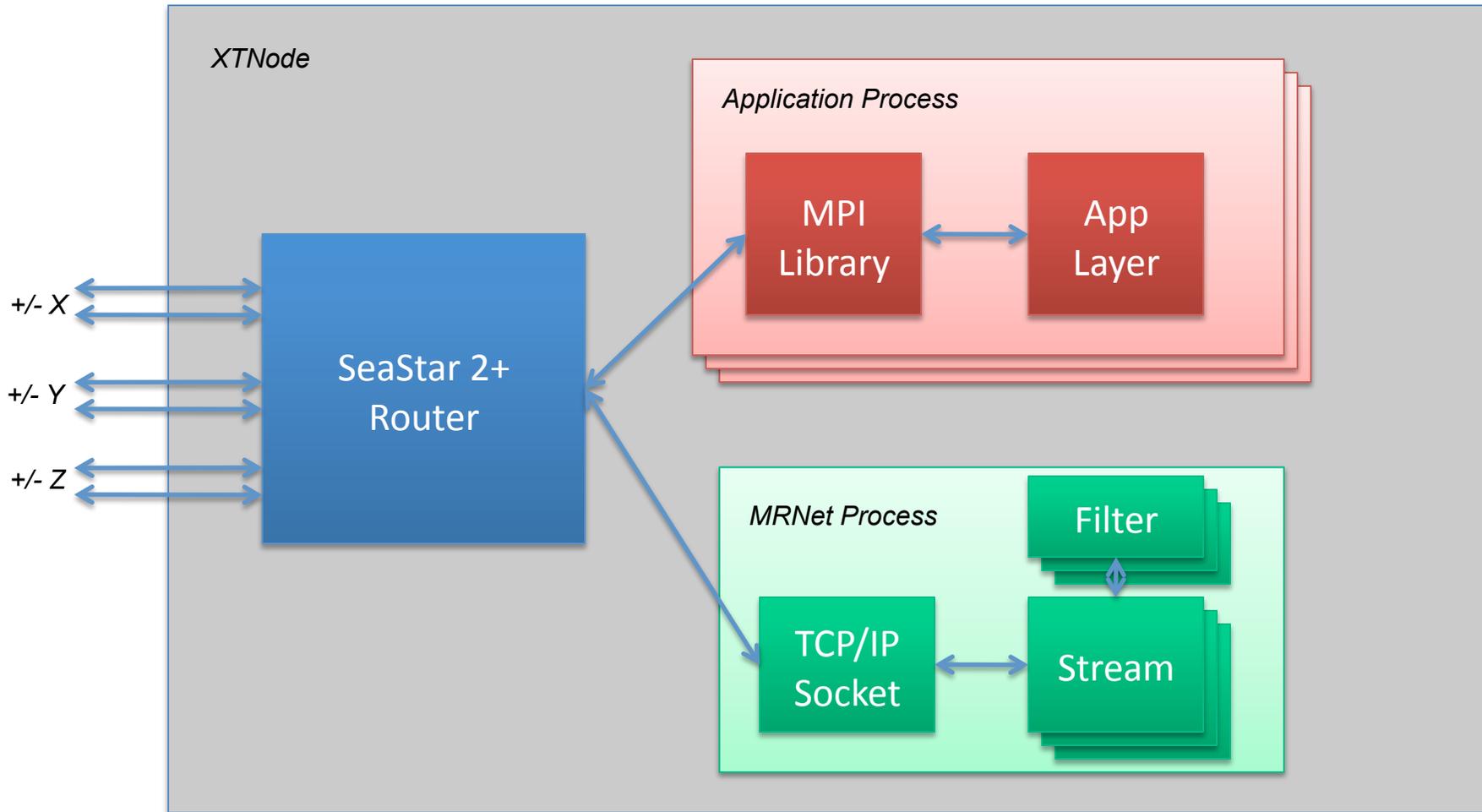
MA-Instrumented MPI Program

System Model

- Node modules connected in 3D torus
- Implemented using OMNeT++ (<http://www.omnetpp.org>)

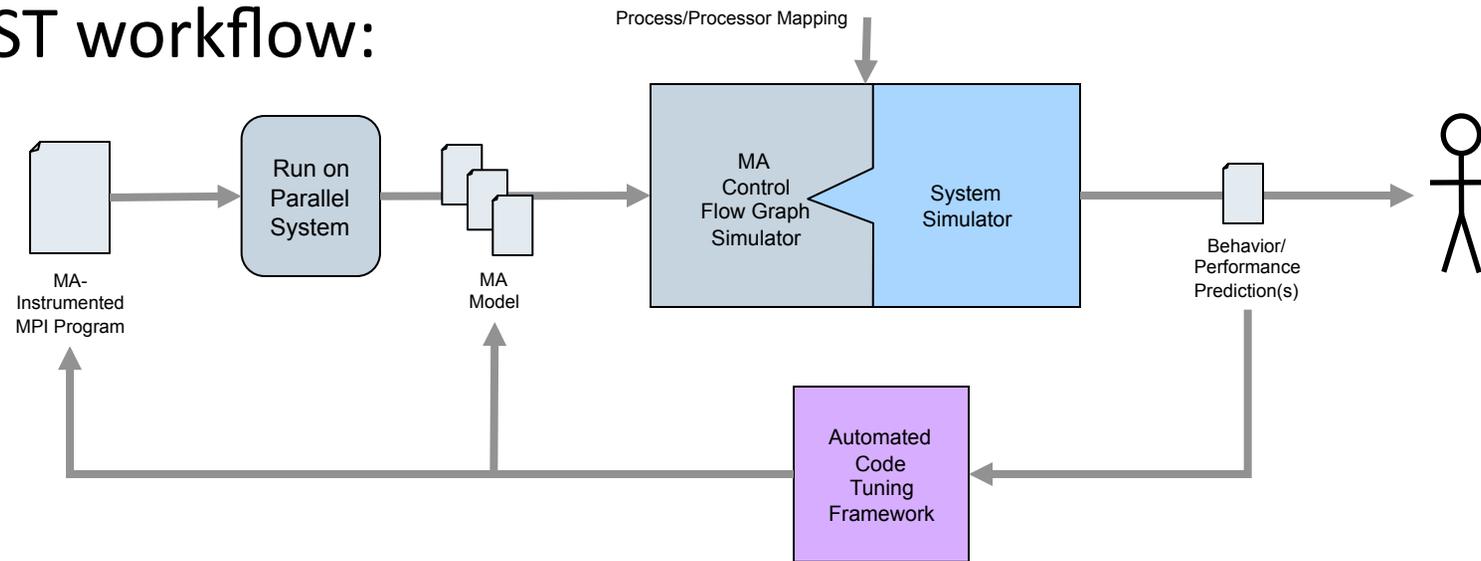


Node model

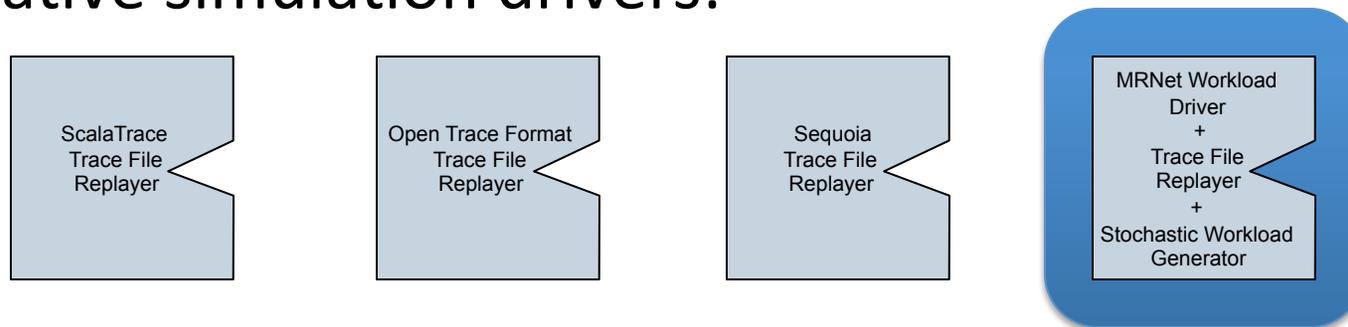


Simulation Flexibility

- MAST workflow:



- Alternative simulation drivers:



Status

- Basic XTNode with SeaStar router is implemented
- Support for simple MPI-based workloads
 - Hardcoded behaviors (e.g., ping pong or ring)
 - OTF and Sequoia trace readers
- Support for TBON processes designed and partially implemented
- Recently ported from OMNeT++ 3.2 to 4.0 – still working out changes in OMNeT++ handling of simulation time
- Expect TBON predictions within this month

Comments and Common Questions

- Simulation framework should be easy to adapt to alternative platforms
 - InfiniBand-based clusters (using Mellanox-provided OMNeT++ InfiniBand model)
 - Nodes with compute accelerators (e.g., Keeneland, OLCF3 with GPUs)
- Why not use analytic models?
 - Derivation of analytic models can be difficult
 - MRNet and application processes in XT mesh/torus network
 - Possibility of varying hop counts between logically connected processes
 - Contention
- Need higher fidelity?
 - Want to use least level of detail needed to provide good predictions
 - Increases simulation scalability, potential for online predictions (e.g., to support dynamic TBON reconfiguration)
- Why not use Structural Simulation Toolkit (SST)?
 - ORNL and Sandia are partners in Institute for Advanced Architectures and Algorithms (IAA)
 - SST was not ready when this work started
 - SST focus on highly detailed simulation of a few nodes

The Scalable Heterogeneous Computing (SHOC) Benchmark Suite

- Large systems with non-traditional compute devices are coming (or already here)
- SHOC benchmark suite tests performance and stability of such systems
 - Initial focus on GPUs and multi-core processors
 - Supports clusters and individual hosts
 - Programs in OpenCL and CUDA
- <http://ft.ornl.gov/doku/shoc/start>



Acknowledgements

- This research is sponsored by the Office of Advanced Scientific Computing Research; U.S. Department of Energy. The work was performed at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC under Contract No. De-AC05-00OR22725.
- This research used resources of the Center for Computational Sciences at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. De-AC05-00OR22725.

Summary

- Predicting TBON performance on Cray XT is highly desirable
 - Matching TBON process topology and placement to tool needs subject to application and system constraints
 - May support online reconfiguration of TBON topology
- Developing simulation-based TBON prediction capability
 - Expect predictions of realistic scenarios soon
 - Easily adaptable to expected future architectures

Questions ??

<http://ft.ornl.gov>

