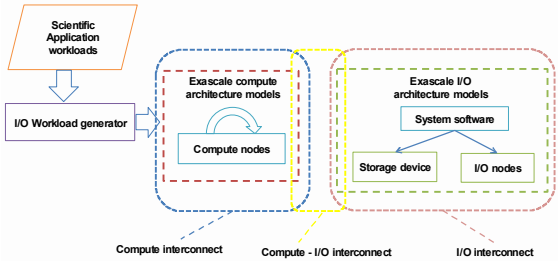
On the role of Parallel Simulation in Extreme-Scale Storage & Network Co-design

Misbah Mubarak (mubarm@cs.rpi.edu), Christopher D. Carothers (chrisc@cs.rpi.edu) - Rensselaer Polytechnic Institute
Robert B. Ross (rross@mcs.anl.gov), Philip Carns (carns@mcs.anl.gov) - Argonne National Laboratory

The CODES PROJECT

Enabling **CO**-**D**esign of **E**xascale **S**torage Systems

The goal is to accelerate exascale storage architecture design via detailed simulation of storage components, software, and the surrounding environment.
Incrementally develop storage simulation capability, validating approaches and components along the way.



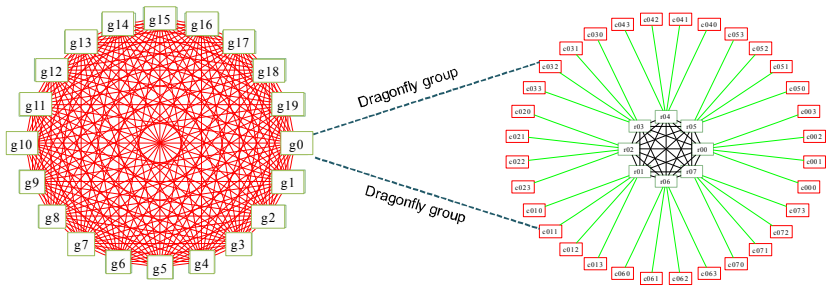
ROSS

Rensselaer **O**ptimistic **S**imulation **S**ystem

ROSS provides Parallel Discrete Event Simulation (PDES) capability for CODES. ROSS has the capability of optimistically scheduling events using reverse computation. It has been shown to process billions of events per second on Sequoia Blue Gene/Q supercomputer.

The CODES simulation infrastructure can be used to track Exascale system development and investigate critical aspects for example exascale storage systems and interconnects.

EXPLORING CANDIDATE EXASCALE INTERCONNECTS

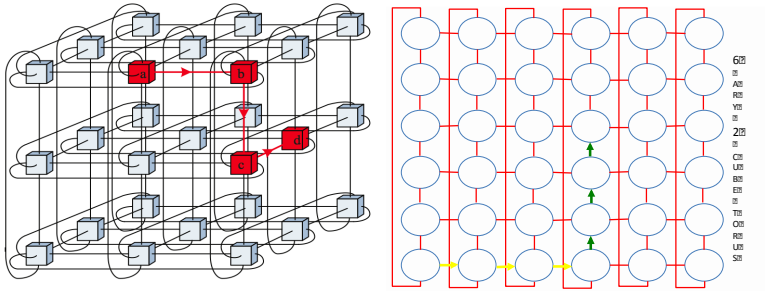


Dragonfly Network Topology: A low diameter, high global bandwidth interconnect.
Pros: Uses high-radix routers to provide very high bisection bandwidth.
Cons: nearest-neighbor communication can become a bottleneck.

A high-bandwidth low-latency interconnect is a critical component of exascale systems.

We need to explore the design space of interconnects to determine a high-performance, low-latency network topology.

Parallel simulation will play an important role to determine a suitable exascale network configuration.



Torus Network Topology: A k-ary n-cube interconnect
Pros: Uses physical locality to produce very high nearest-neighbor throughput
Cons: Limited bisection bandwidth due to higher hop count

MILLION-NODE TORUS AND DRAGONFLY NETWORK MODELS

We have used ROSS discrete-event simulator to model and simulate million-node dragonfly and torus networks at a flit-level detail using well-known HPC traffic patterns.

HIGH FIDELITY & HIGHLY ACCURATE

Torus model accurately reflects static routing behavior of the Blue Gene architecture. Dragonfly model reflects the behavior of booksim, the simulator used to propose the dragonfly topology. Detailed flit-level simulations can execute up to 1 billion events per second on Mira Blue Gene/Q.

Fig: Validation of ROSS torus model with Blue Gene architecture using mpptest performance benchmark.

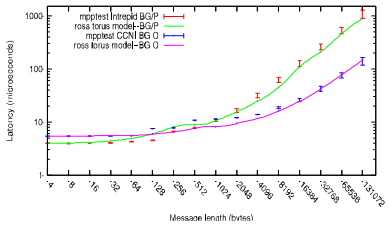
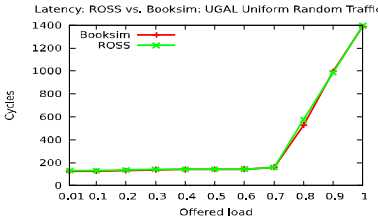


Fig: Validation of ROSS dragonfly model with the booksim simulation frameowork.



CONFIGURABLE

We have modeled 5-D, 7-D and 9-D configurations of a torus network to observe the effect of torus dimensionality on network performance under heavy network traffic.

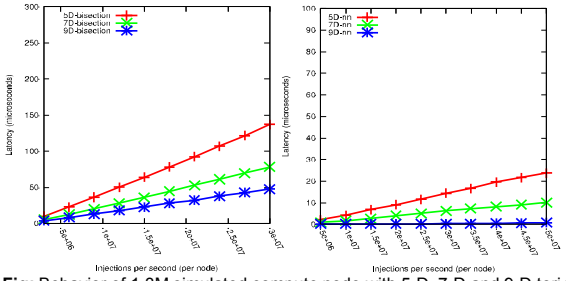
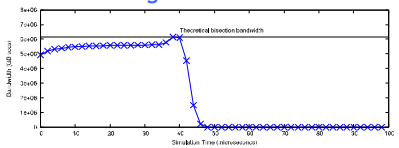


Fig: Behavior of 1.3M simulated compute node with 5-D, 7-D and 9-D tori with diagonal (bisection) and nearest-neighbor traffic patterns on 1 rack of Mira BG/Q.

We have modeled a 1.3M nodes dragonfly network and identified its best and worst-case configurations

Fig: The router-router link of a 1.3M simulated compute node dragonfly model is congested with network traffic which yields low network bandwidth.



HIGHLY EFFICIENT

The exascale size simulations of torus and dragonfly networks run efficiently on Argonne's Mira Blue Gene/Q system.

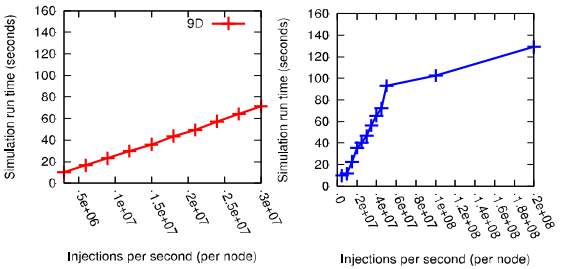


Fig: Simulation run time of ROSS 9-D torus and dragonfly models with 1.3 M compute nodes on 65,536 MPI tasks of Mira BG/Q.

The work was supported by the Office of Advanced Scientific Computer Research (ASCR) under contract DE-AC02-06CH11357. The research used resources from Argonne Leadership Computing Facility (ALCF).