Lessons From Analyzing Fan-In Communications

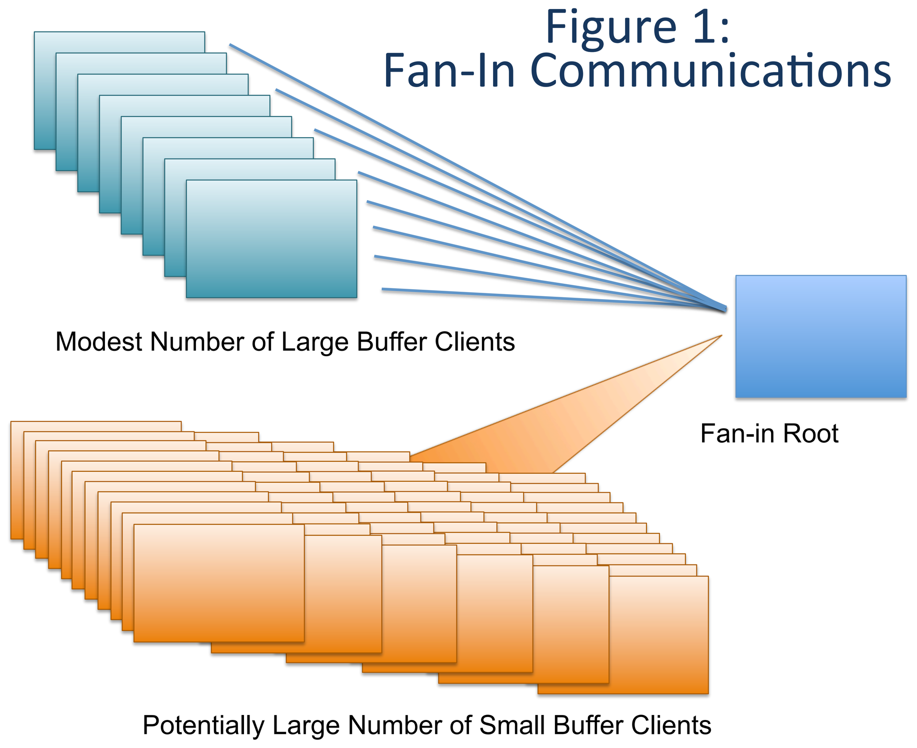
Unconstrained Performance @ Exascale

Principal Investigators: Terry Jones ORNL, Bradley Settlemyer; ORNL

Motivation

- Fan-in communications are a key performance issue for many parallel applications including both scientific applications and system software components. Our work is part of an effort to address the challenging problem of file system quality of service.
- By their nature, fan-in communications form "hot spots" that present significant challenges to interconnects.

Figure 1:
Fan-In Communications



Trend towards more cores underscores the need for better understanding of hotspot issues. fan-in communication provides an important design point.

- We explore multiple factors from the large design space to understand when the various choices matter, and the magnitude of their impact. The study also clarifies certain limitations surrounding the use of latency mitigation techniques including "burst buffers."

Approach

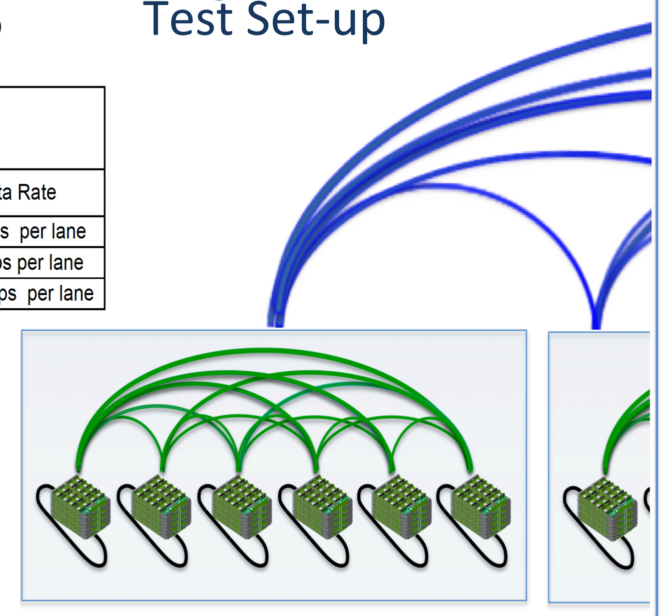
- We are able to assess the impact of multiple communication factors for Extreme Scale Environments by conducting multiple measurements of a basic fan-in scenario in which a variable number of clients exchange messages with one root node. Specifically, we conducted tests on a large Cray Aries-interconnect based supercomputer and a large Cray Gemini-interconnect based supercomputer) each under varying conditions to study the impact of fan-in congestion
- Scenarios included multiple high-priority clients exchanging large buffers, and multiple low-priority clients exchanging small buffers.
- Our study included both synchronous and asynchronous patterns.

Figure 2:
Test Set-up

Blue: 1 Machine= n Groups
Green: 1 Group = 6 Chassis (384 nodes)
Black: 1 Chassis (64 nodes)

ARIES Speeds
gbps = Giga bits per second

Aries Component	Link Type	Data Rate
Black	PCI Express Gen3 x16	16 gbps per lane
Green	Electrical cables	14 gbps per lane
Blue	Optical Cable	12.5 gbps per lane



For Eos: Blue=240 links using 60 optical cables

Results

- Our results present fan-in performance in terms of aggregate bandwidth while controlling the degree of congestion and several other key attributes.
- Each graph shows aggregate bandwidth as measured at the fan-in root on the Y-axis versus the number of large-buffer clients on the X-axis (higher numbers on Y-axis is better). In each test, large-buffer is defined as 4MB, small buffer is defined as 4K.

Figure 3: Fan-In With No Chatter Clients

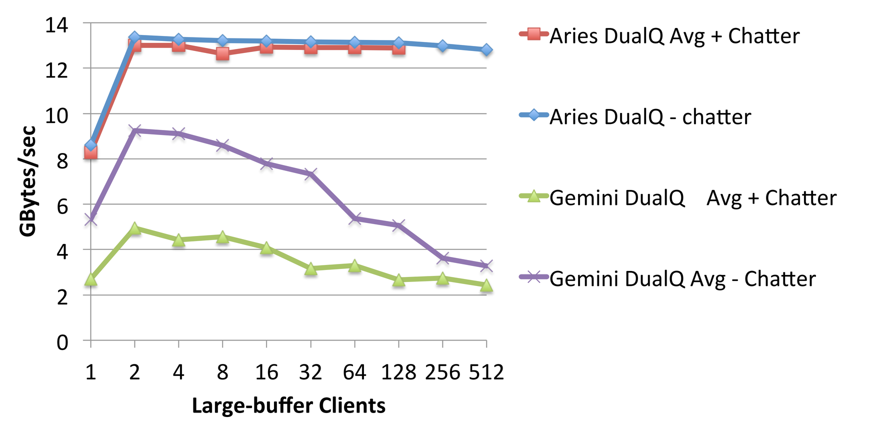


Figure 4: Variability

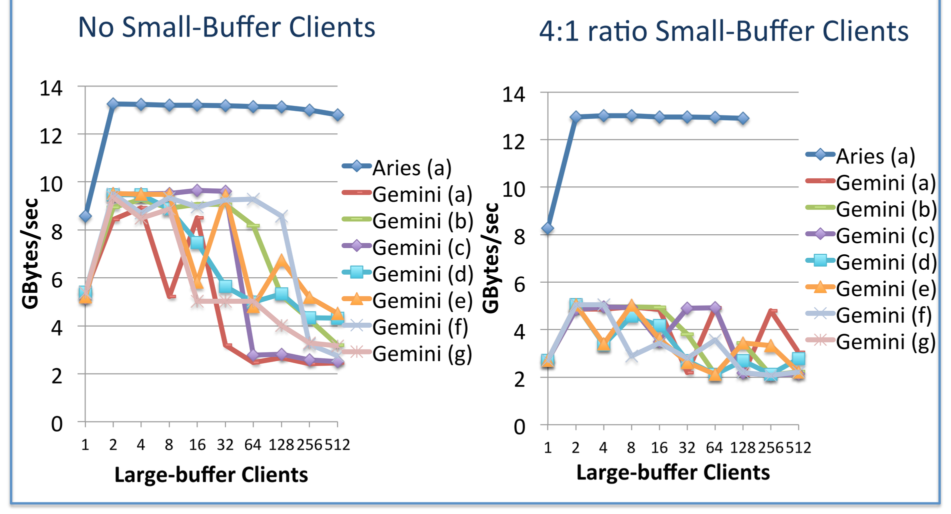


Figure 5: The Impact of Asynch Unexpected Messages

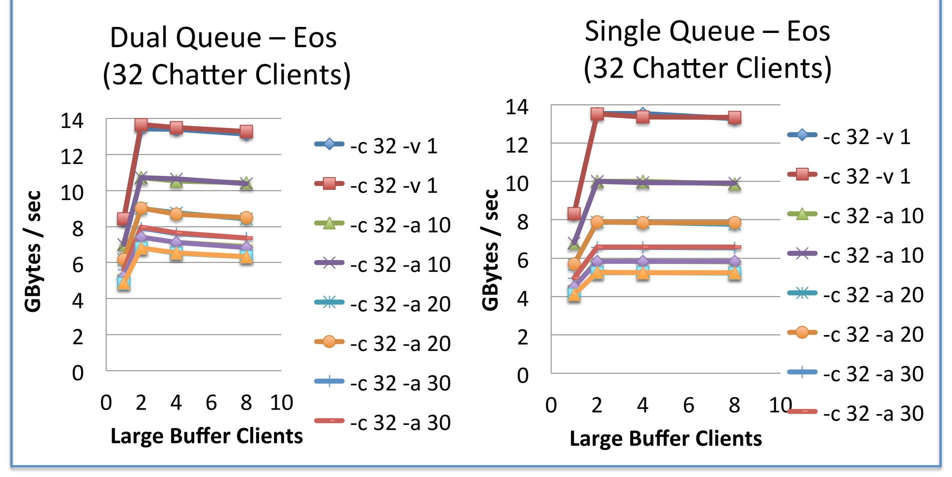
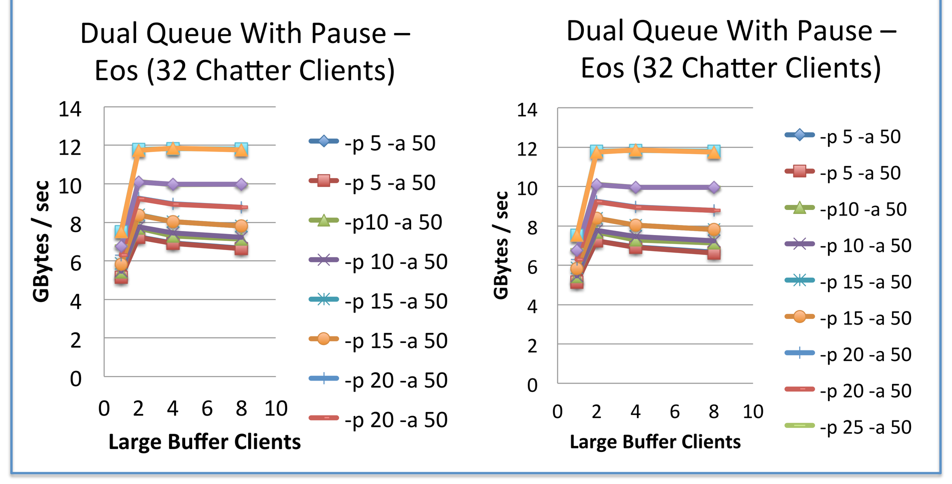


Figure 6: The Effectiveness of Reservation Schemes



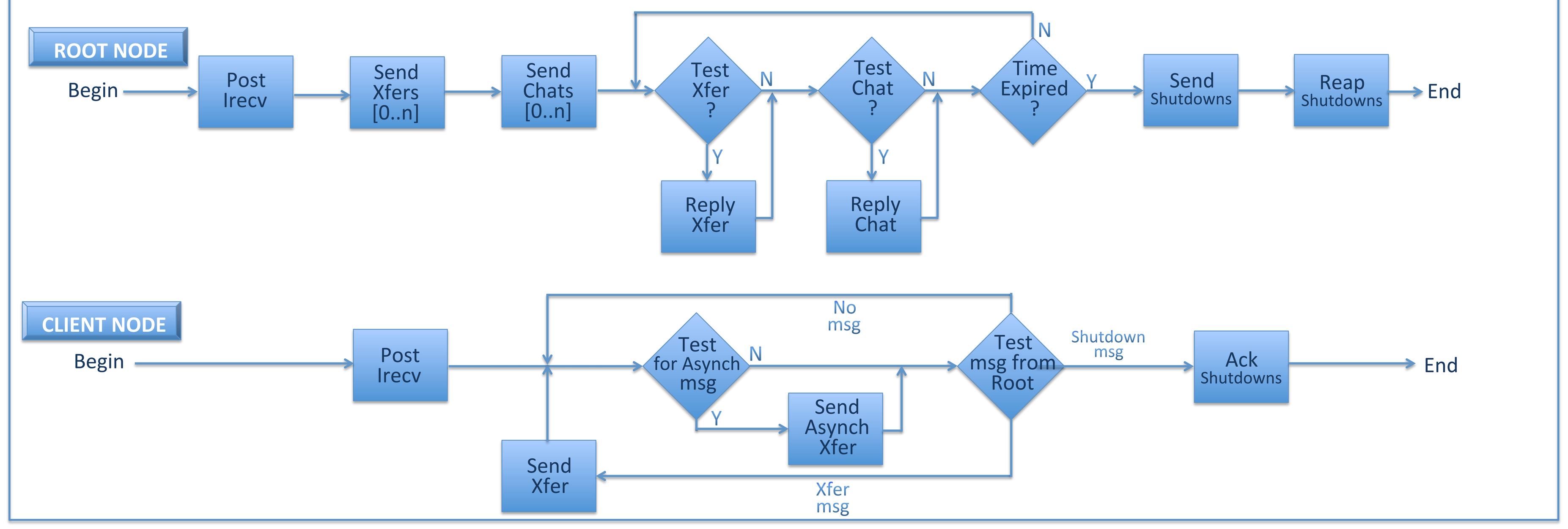
Lessons Learned:

- ✓ Once a threshold of unexpected messages is reached, performance for both Aries and Gemini fall linearly with each additional message. This condition is much more prevalent under asynchronous patterns.
- ✓ Neither Aries nor Gemini are able to reach their peak bandwidth with only one large-buffer client (i.e., two or more clients are required to saturate a servers bandwidth capacity).
- ✓ Aries is able to achieve slightly more than 13 GB/sec bandwidth at the root; Gemini is able to achieve slightly more than 9 GB/s.
- ✓ Aries bandwidth does not vary considerably from run to run or drop considerably under increased contention (more client nodes), the depth of the non-blocking receive queue, the rank-to-node mapping, or the presence or absence of wildcard matching for either message source or message tag.
- ✓ Gemini drops between 50% and 20% under contention from small-buffer clients, and exhibits a constant bandwidth of around 5.1 GB/sec regardless of the number of clients when the message source and tag are restricted.

Novelty

- Using a newly developed tool, Reservation, we are able to test user level reservation schemes to explore file system quality of service. The tool allows one to arbitrarily set key parameters such as the scale of the fan-in tree, the level of asynchrony (i.e., messages in flight), the number of clients and the size(s) of their messages, and details about request queues and fairness properties.

Figure 7: Reservation Tool Logic (simplified)



U.S. DEPARTMENT OF ENERGY

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