Understanding Network Contention on Blue Gene Supercomputers

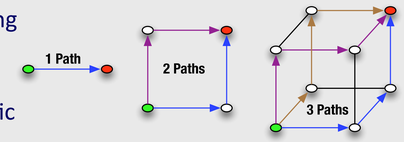
Nikhil Jain, Abhinav Bhatele, Todd Gamblin, Martin Schulz, Laxmikant V. Kale

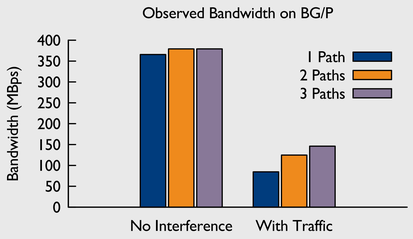
Topology aware task mapping can improve performance of parallel applications. It is widely accepted that mapping can optimize latency or bandwidth and minimize network contention. However, the correlation between different mappings, routing schemes, network contention and performance improvements is not well understood. We present preliminary work on an accurate understanding of the software stack and hardware design of the torus on Blue Gene machines, their effect on performance, and usefulness of hardware counters in correlating mapping & performance. We demonstrate up to 12% improvement in observed bandwidth by altering the software stack based on our understanding of these issues.

Critical factors affecting messaging

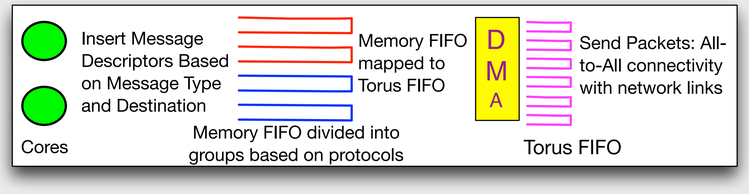
Experiment: measure ping-pong bandwidth

- Mode 1: no interference
- Mode 2: with all-to-all traffic in background





- Without interference, extra paths do not increase observed bandwidth – Why?
- With interference, extra paths help – Why?

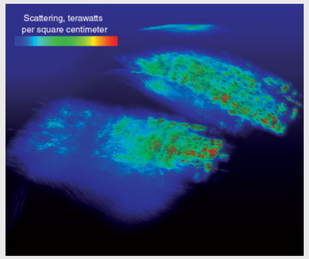


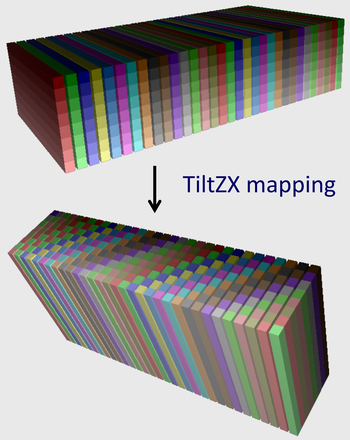
Message Injection

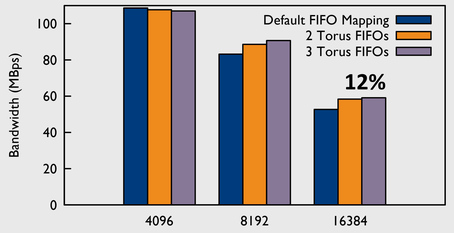
Impact on application performance

pF3D is a multi-physics code used to study laser plasma-interactions at NIF, LLNL. It has two communication phases:

- All-to-alls for FFT
- Pairwise exchange



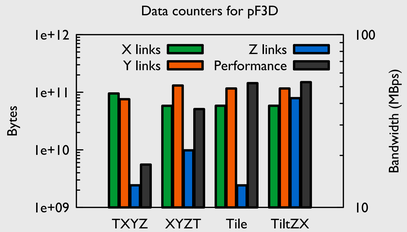




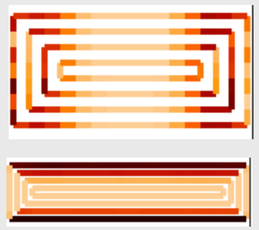
- Most of the gain is found in pairwise exchange phase
- Improvements are over the fine tuned task mapping

Network counters versus performance

- Amount of traffic passing through links does not correlate with actual performance
- Do we need to look at other counters?



2D projection of traffic flowing on deterministic channels for TXYZ and TiltZX mappings – darker shades indicate more traffic

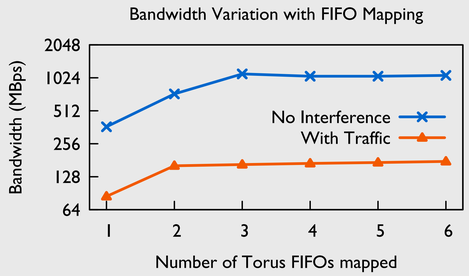


Dynamic and deterministic channels

Histograms showing ratio of traffic on dynamic channels to that on deterministic channels – ***concentration towards right is better***

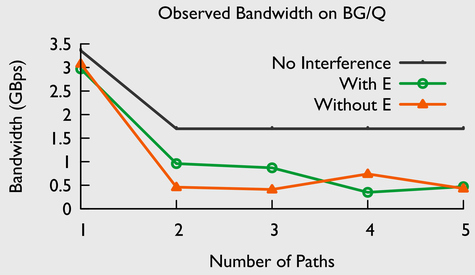
Can FIFO mapping improve performance?

- Default is 1 memory FIFO to 1 torus FIFO
- Change mapping in DCMF code (10 lines)
- Bottleneck may shift from serialization within a message to serialization between messages



Experiments on Blue Gene/Q

- The fifth (E) dimension (of size 2) behaves in a unique manner
- Instead of FIFO mapping, BG/Q introduces message mapping – finer control over routing of messages
- More freedom to map messages to queues, but need to strike a balance between routing on multiple paths and contention for links



Contact Author – Nikhil Jain – nikhil@illinois.edu

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-POST-568293