Avoiding hot-spots on two-level direct networks

Presented by: Alexandros Papados (AMSC)

University of Maryland, College Park:
CMSC 714: High Performance Computing

April 8,2021

Table of Contents

Table of Contents

Introduction

- This paper explores topology aware mappings of different communication patterns to the physical topology to identify cases that minimize link utilization
- Analyzes the trade-offs between using direct and indirect routing with different mappings
- Simulations are used to study communication and overall performance of applications since there are no installations of two-level direct networks
- Raises interesting issues regarding the choice of job scheduling, routing and mapping for future machines

Table of Contents

Major Contributions

- This paper has the first analysis of congestion on a two-level direct topology due to routing and mapping choices.
- Presents several solutions for avoiding hot-spots on such networks.
- The paper presents the largest packet-level detailed network simulations done so far (for 307,200 cores) for several communication patterns.
- Presents several mappings for 2D, 4D and multicast patterns and compare their performance when coupled with direct and indirect routing on the PERCS network.

Table of Contents

The Percs Topology

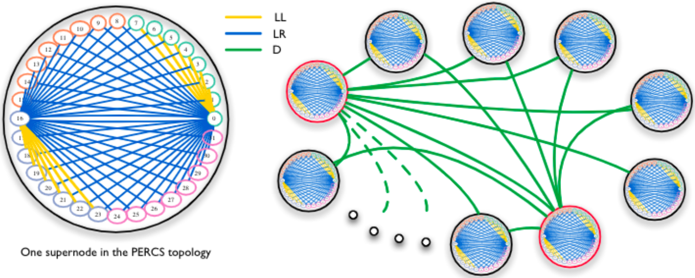

Figure 1: The PERCS network – the left figure shows all to all connections within a supernode (connections originating from only two nodes, 0 and 16, are shown to keep the diagram simple). The right figure shows second-level all to all connections across supernodes (again D links originating from only two supernodes, colored in red, are shown).

The Percs Topology

- The PERCS interconnect topology is a fully connected two-tier network
- Within the large circle, a small circle represents a quad chip module (QCM) which consists of four 8- core Power7 chips.
- Eight nodes in one color in each quadrant constitute a drawer.
- Each node has a hub/switch which has three types of links originating from it - LL, LR and D links.
- There are seven LL links (24 GB/s) that connect a node to seven other nodes in the same drawer.
- In addition, there are 24 LR links (5 GB/s) that connect a node to the remaining 24 nodes of the supernode.

The Percs Topology

- For a system with n supernodes, the number of D links is $(n \times (n-1))$. There are $(32 \times 31 \times \mathrm{n})$ LL and LR links in total. Hence, there are $(992/(n-1))$ first tiers links for every second tier link as shown in Figure 2.

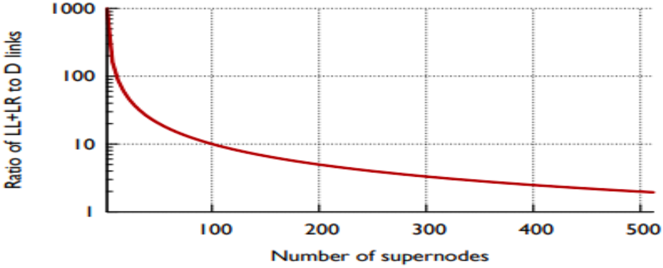

Figure 2: The number of D links reduces significantly compared to that of LL and LR links as one uses fewer and fewer supernodes in the PERCS topology.

Table of Contents

Topology aware mapping

- Topology aware mapping of MPI tasks to physical cores/nodes on a machine can minimize contention and impact application performance
- **Types of Mappings**
  - Default Mapping (DEF)
  - Blocked Nodes Mapping (BNM)
  - Blocked Drawers Mapping (BDM):
  - Blocked Supernodes Mapping (BSM)
  - Random Nodes Mapping (RNM)
  - Random Drawers Mapping (RDM)

Table of Contents

Simulations for 64 Node

- We now present simulation results for communication patterns summarized in Table 2.
- Simulations were done for 64 supernodes.

Table II

Communication Pattern	Number of Supernodes	Number of Elements	Number of Messages	Message Size (KB)	Sequential Computation (ms)
2D 5-point Stencil	64	8192×8192	4	64	479
4D 9-point Stencil	64	$64 \times 64 \times 64 \times 64$	8	2048	224
Multicast Pattern	64	–	14	1024	–
4D 9-point Stencil	300	$64 \times 32 \times 64 \times 32$	8	1024	50

Table 2: Details of the experimental setup for different communication patterns and different number of supernodes

Mapping a 2D 5-point Stencil

The data array for this 2D Stencil is 2097152 $\times$ 2097152 and each MPI task is given a sub-domain of 8192$\times$8192 elements. This gives us a logical 2D array of MPI tasks of dimensions 256 $\times$ 256 which is to be mapped to 65, 536 cores (64 supernodes)

Mapping	Node	Drawer	Supernode
DEF	32×1	256×1	256×4
BNM	8×4	64×4	256×4
BDM	8×4	16×16	64×16
BSM	8×4	16×16	32×32

Table 3: Dimensions of blocks at different levels (node, drawer and supernode) for different mappings of 2D Stencil

Mapping a 2D 5-point Stencil

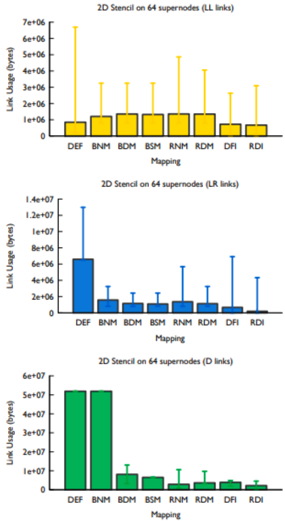

Figure 5: Average number of bytes sent over LL, LR and D links for 2D Stencil on 64 supernodes

Table of Contents

Conclusions

- Multi-level direct networks have emerged as a new technology to connect a large number of processing elements together.
- Default MPI rank-ordered mapping with direct routing on such networks leads to significant hot-spots, even for simple two and four dimensional near-neighbor communication patterns.
- Discusses techniques and analyzes various choices for congestion control on these networks.
- Used detailed packet-level network simulations for up to three hundred thousand MPI tasks and three different communication patterns to compare various mappings – default mapping, blocked mapping to nodes, drawers, or supernodes and mapping to random nodes and drawers.

Table of Contents

Reference

Bhatele, Abhinav, et al. "Avoiding Hot-Spots on Two-Level Direct Networks." Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis on - SC '11, 2011, doi:10.1145/2063384.2063486.