Scalable Fault-tolerant Interconnection Networks for Large-scale Computing Systems

Jose Carlos Sancho
Ana Jokanovic and Jesus Labarta

Barcelona Supercomputing Center, Spain
Outline

- HPC network interconnects trends
- Future network failure expectations
- Handling errors in today’s HPC networks
- Proposed scalable fault-tolerant approaches
  - Permanent or hardware failures
  - Transient failures or bit errors
- Conclusions
Optical interconnects are preferred over copper

- **Size**: Lower cable size & weight, Smaller connector size
- **Attenuation**: Longer distance, more power-efficient transmission at higher BW*distance
- **Isolation**: Lower EMI, less crosstalk, more reliable

**Active Optical Cable**
- Opto/Electronic conversions (CMOS integration)
- Bandwidth 40 Gbps (4 x 10 Gbps)
- $< 10^{-15}$ Bit Error Rate (temporal or soft failures)
- $5 \times 10^6$ hours Mean time between permanent failures
HPC networks size

Top10 historical data

Source: Peter M. Kogge and Timothy J. Dysart, “Using the TOP500 to Trace and Project Technology and Architecture Trends”, SC11
HPC network bandwidth is increasing

Source:
The Energy issue is well understood

- Systems are constrained by power, exascale < 20MW

Data movement is the key to save energy

- Up to 200x more energy needed to transport a bit from a nearest neighbor chip than to operate on it:
  - Energy needed for a floating point operation (0.1-0.05 pJ/bit)
  - Energy needed for (electronic) data-transport on card (2-10 pJ/bit)

Sources:
DARPA/IPTO study, by Peter Kogge, et. al. available on http://www.nd.edu/~kogge/reports.html
Optical transceiver power

- Power scales with data rate

Source:
Low power increases BER

Power (in dBm) = $10 \log_{10} \frac{\text{Power}}{1\text{mW}}$

1 mW = 0 dBm
1 µW = -30 dBm

Source:
Power wall implications

Trading off power with BER @ high data rates

Low failures
Low power
Handling errors in InfiniBand (1)

Hardware failures

Fault link:
- Discard packet
- Inform SMA

Subnet Manager (SMA)
- Re-program routing tables at S8
- Inform A0
Handling errors in InfiniBand (2)

Bit errors

Checking CRC
Discard faulty packet

Switch receiving packet

EBPD | header

Switch receiving packet

Discard faulty packet

Compute source node

Packet retransmission

VCRC | CRC | data payload | header

NACK

Compute destination node

Too much overhead
- Notification
- Re-transmission
Scalable fault-tolerant approaches

- Longer distances between end nodes prevent efficient error handling.

Handle it locally!
Local hardware failure recovery

Fault link:
- Re-route packet
- Inform SMA

Open issues
- Deadlock?
- Who to re-route packet?

Subnet Manager (SMA)
- Re-program routing tables at S8
Local bit error recovery

- Use Error Correcting Codes (ECC)
- Populate packets with ECCs

Open issues
- Which ECC?
- Communication protocol?

Local failure recovery

Switch receiving packet

Packet free of errors

Switch receiving packet

Compute source node

NACK

Compute destination node
Conclusions

- Bit errors and hardware failures may happen more often in future computing systems
  - The number of network components is growing
  - Power wall increases BER

- Current networks are still relaying on end-node recovery which adds too much overhead

- Locally handling errors is promising but still requires efficient techniques to be explored which may be different depending on the type of failure
Acknowledgements

- Spanish Ministry of Science and Innovation