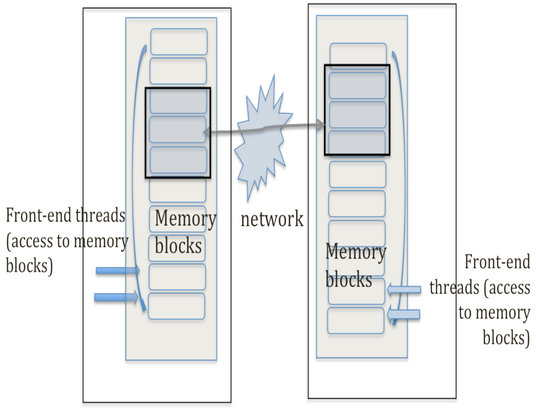
MemzNet: Memory-Mapped Zero-copy Network Channel for Moving Large Datasets over 100Gbps Networks

Mehmet Balman (mbalman@lbl.gov)

Computational Research Division, Lawrence Berkeley National Laboratory

Collaborators: Eric Pouyoul, Yushu Yao, E. Wes Bethel, Burlen Loring, Prabhat, John Shalf, Alex Sim, Arie Shoshani, Dean N. Williams, Brian L. Tierney

Memory-mapped Network Channel Framework



memory caches are logically mapped between client and server

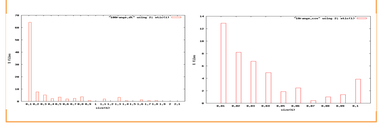
Increasing the bandwidth is not sufficient by itself; we need careful evaluation of future high-bandwidth networks from the applications' perspective. We require enhancements in current middleware tools to take advantage of future networking frameworks. To improve performance and efficiency, we develop an experimental prototype, called MemzNet: Memory-mapped Zero-copy Network Channel, which uses a block-based data movement method in moving large scientific datasets. We have implemented MemzNet that takes the approach of aggregating files into blocks and providing dynamic data channel management. We present our initial results in 100Gbps networks.

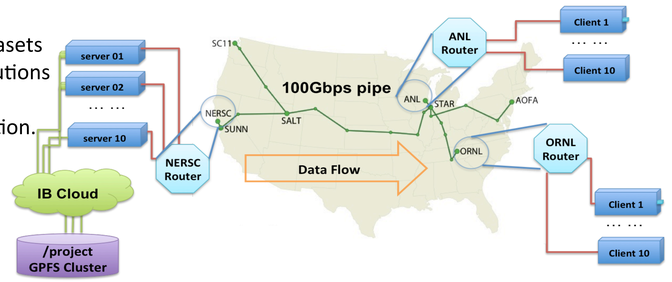
SC11 100Gbps Demo Configuration

Climate Data-file characteristics

- Many small files
- One of the fastest growing scientific datasets
- Distributed among many research institutions around the world
- Requires high-performance data replication.

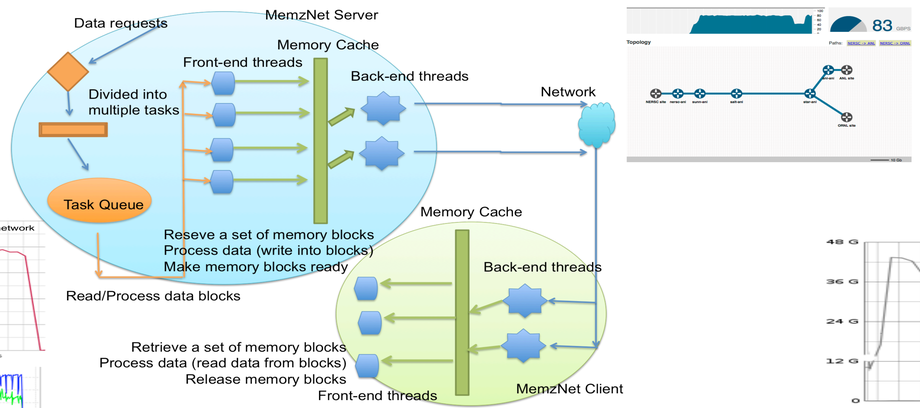
File size distribution in IPCC Fourth Assessment Report (AR4) phase 3, the Coupled Model Intercomparison Project (CMIP3)





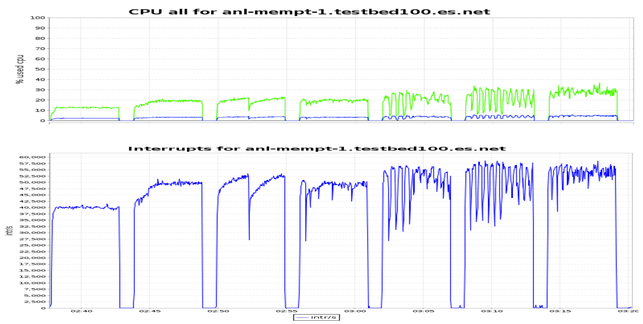
- Many TCP sockets oversubscribe the network and cause performance degradation.
- Host system performance could easily be the bottleneck.

Moving Climate Files Efficiently

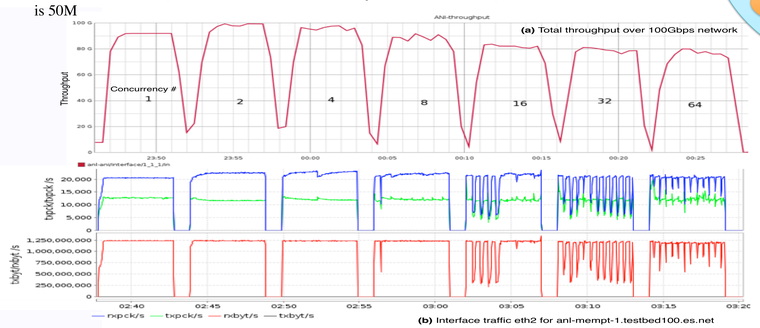


Measurement in ANI 100Gbps Testbed

3 hosts, each connected with 4 10Gbps NICs to 100Gbps router

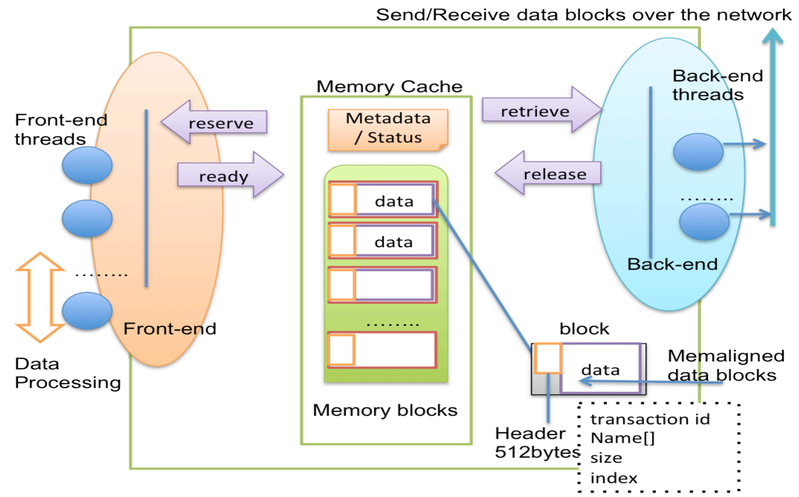


ANI testbed 100Gbps (10x10NICs, three hosts): CPU/Interrupts vs the number of concurrent transfers [1, 2, 4, 8, 16, 32 64 concurrent jobs - 5min intervals], TCP buffer size is 50M



(a) total throughput vs. the number of concurrent memory-to-memory transfers, (b) interface traffic, packages per second (blue) and bytes per second, over a single NIC with different number of concurrent transfers. Each peak represents a different test; 1, 2, 4, 8, 16, 32, 64 concurrent streams per job were initiated for 5min intervals

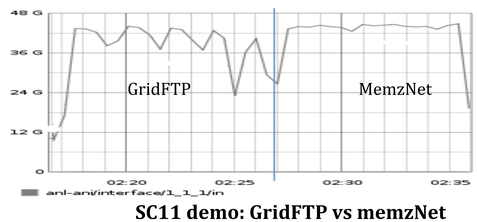
MemzNet's Architecture for data streaming



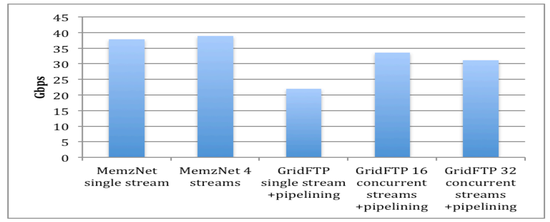
Features

- Data files are aggregated and divided into simple blocks. Blocks are tagged and streamed over the network. Each data block's tag includes information about the content inside.
- Decouples disk and network IO operations; so, read/write threads can work independently.
- Implements a memory cache managements system that is accessed in blocks. These memory blocks are logically mapped to the memory cache that resides in the remote site.
- The synchronization of the memory cache is accomplished based on the tag header. Application processes interact with the memory blocks. Enables out-of-order and asynchronous send receive
- MemzNet is is not file-centric. Bookkeeping information is embedded inside each block. Can increase/decrease the number of parallel streams without closing and reopening the data channel.

Performance



SC11 demo: GridFTP vs memzNet



ANI Tetbed: Throughput comparison

Thanks: **Special Thanks** Peter Nugent, Zarija Lukic , Patrick Dorn, Evangelos Chaniotakis, John Christman, Chin Guok, Chris Tracy, Lauren Rotman, Jason Lee, Shane Canon, Tina Declerck, Cary Whitney, Ed Holohan, Adam Scovel, Linda Winkler, Jason Hill, Doug Fuller, Susan Hicks, Hank Childs, Mark Howison, Aaron Thomas, John Dugan, Gopal Vaswani

Acknowledgements: This work was supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This research used resources of the ESnet Advanced Network Initiative (ANI) Testbed, which is supported by the Office of Science of the U.S. Department of Energy under the contract above, funded through the The American Recovery and Reinvestment Act of 2009

References

- Mehmet Balman et al., ***Experiences with 100Gbps Network Applications***. In *Proceedings of the fifth international workshop on Data-Intensive Distributed Computing*, in conjunction with the ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC'12), June 2012.
- Mehmet Balman, **Streaming Exa-scale data over 100Gbps Networks**, IEEE Computing Now, Oct 2012.