RRAIM: Remote Redundant Array of Inexpensive Memory

Roei Tell, Peter Izsak, Aidan Shribman, Benoit Hudzia
SAP Research Israel, SAP Research CEC Belfast
{roei.tell, peter.izsak, aidan.shribman, benoit.hudzia}@sap.com

Key Concept: Transparent Remote Memory Aggregation

Motivation: extending application/VM memory beyond physical capabilities, without performance hit of local swap.

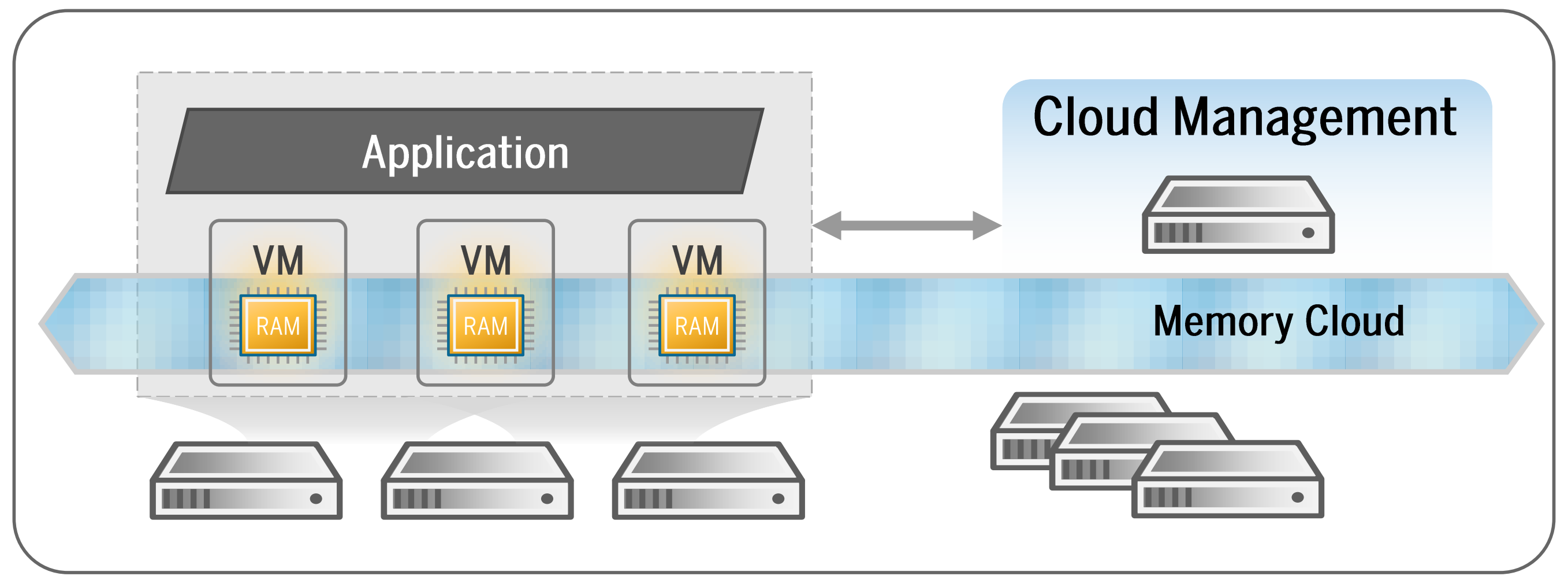
- **Transparent, partially fault-tolerant** remote memory aggregation.
- Leveraging on **low-latency inter-connects** RDMA capabilities.
- Full integration with MMU – **unmodified applications** and VMs.
- Commodity hardware, integrating with existing open-source technologies.

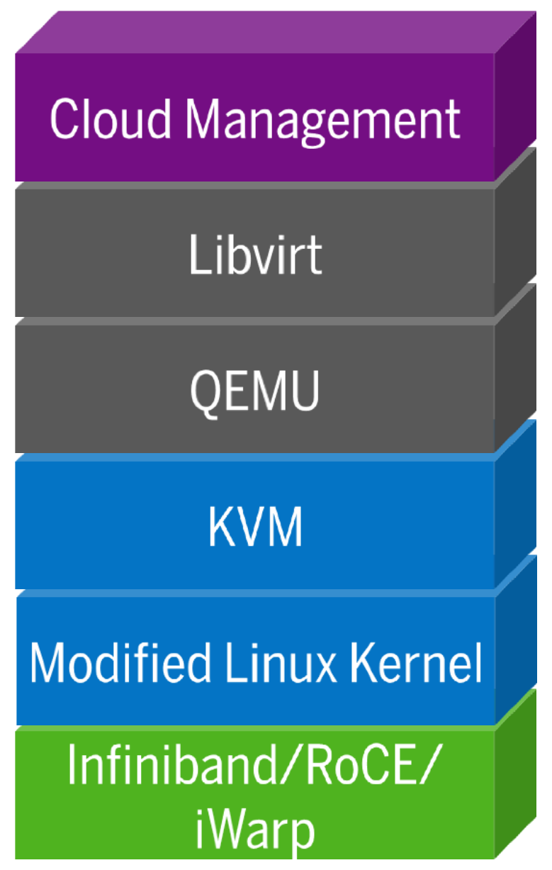
Seamless failover

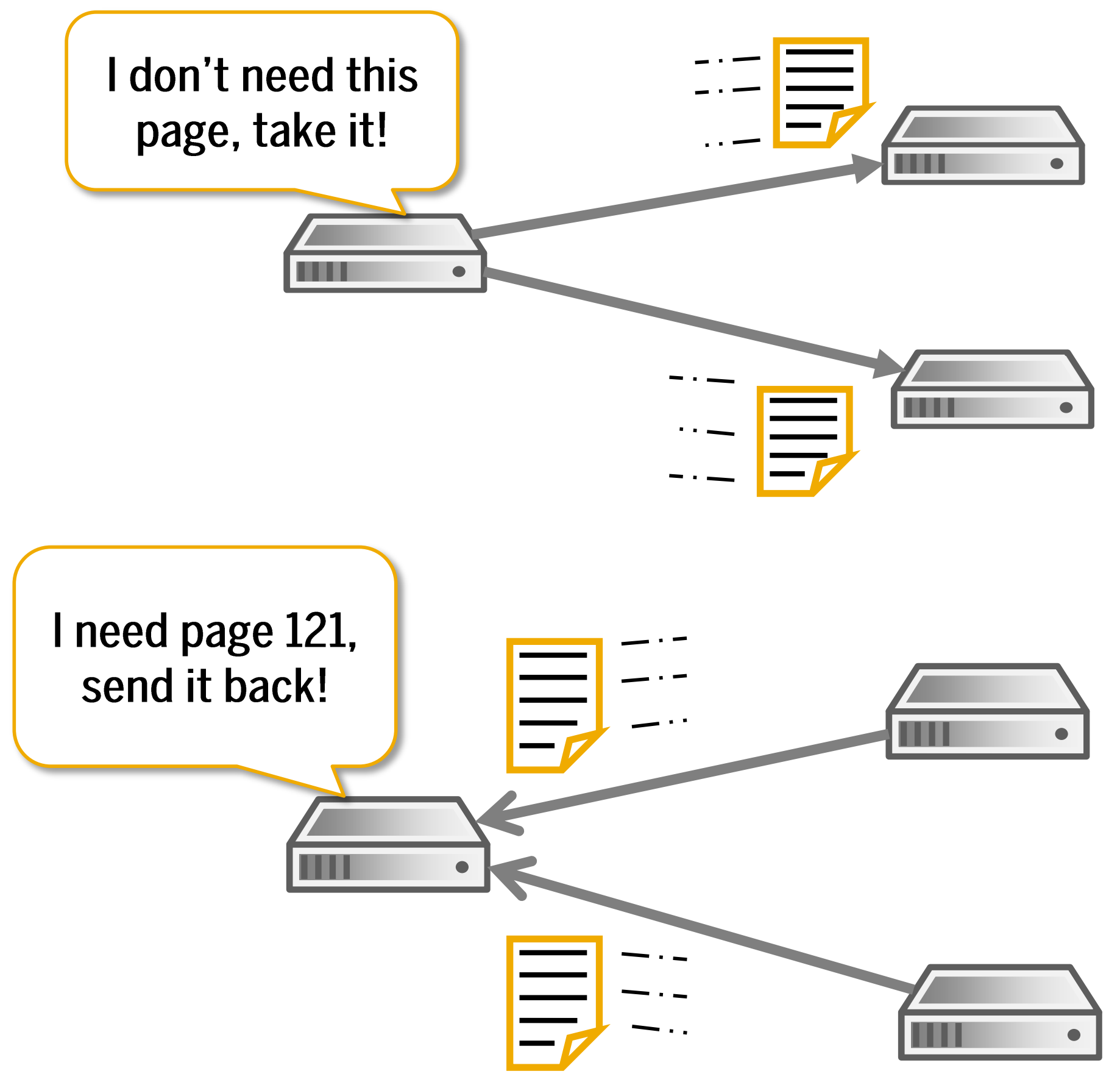
- RAID-1 like schema for remote nodes, page granularity.
- Failure in one remote node does not delay page fault resolves.
- A failed machine is brought back in a linear-time process.

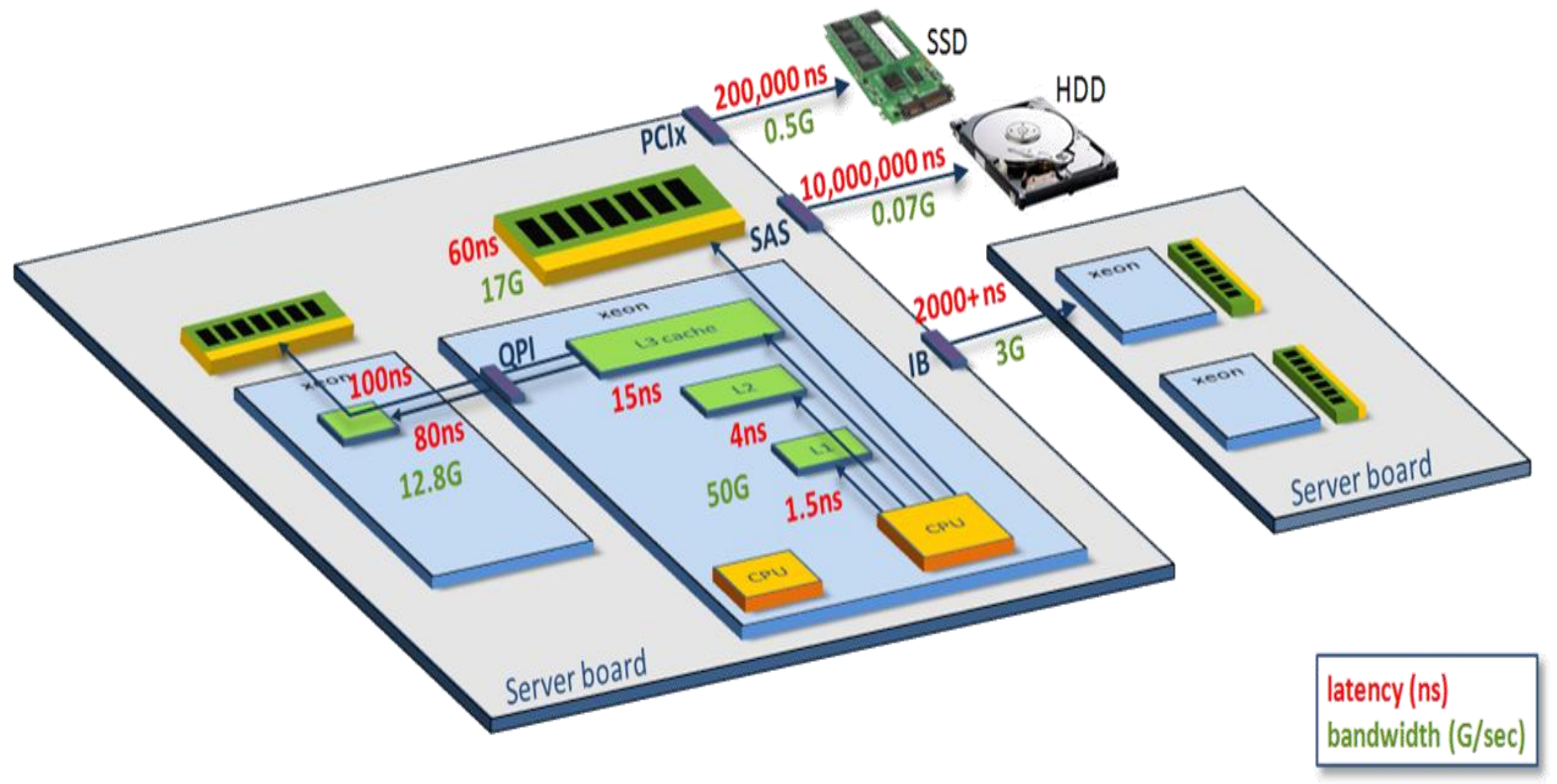
Architecture: Cloud Management, IaaS

- Core implementation in latest Linux Kernel and QEMU codebase.
- Cloud Management solution for an RRAIM Cluster as an IaaS.
- Seamless integration for enterprise apps on Virtual Clusters.
- OFA Verbs interface: compatible with most RDMA implementations.

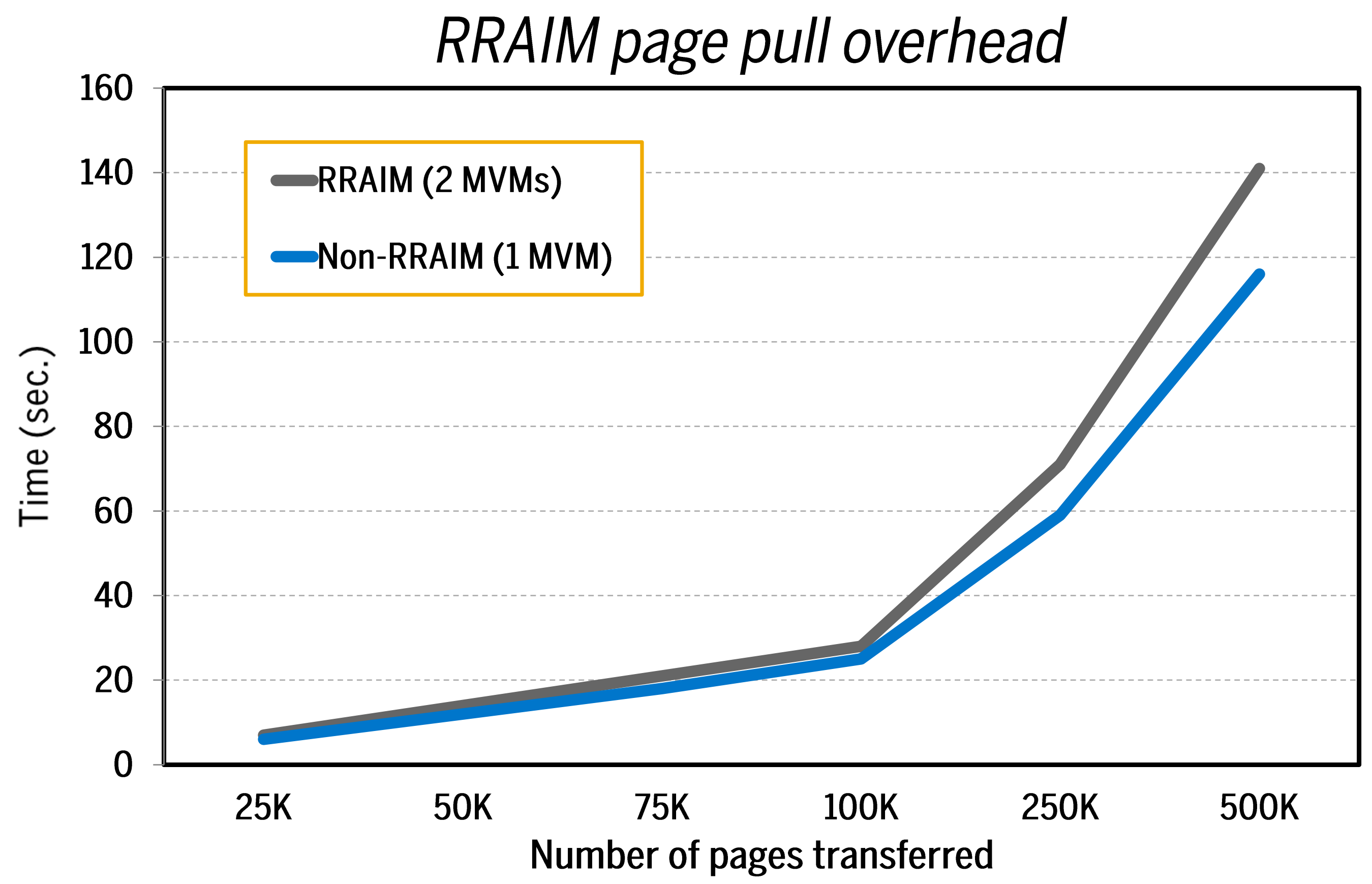





I don't need this page, take it!
I need page 121, send it back!



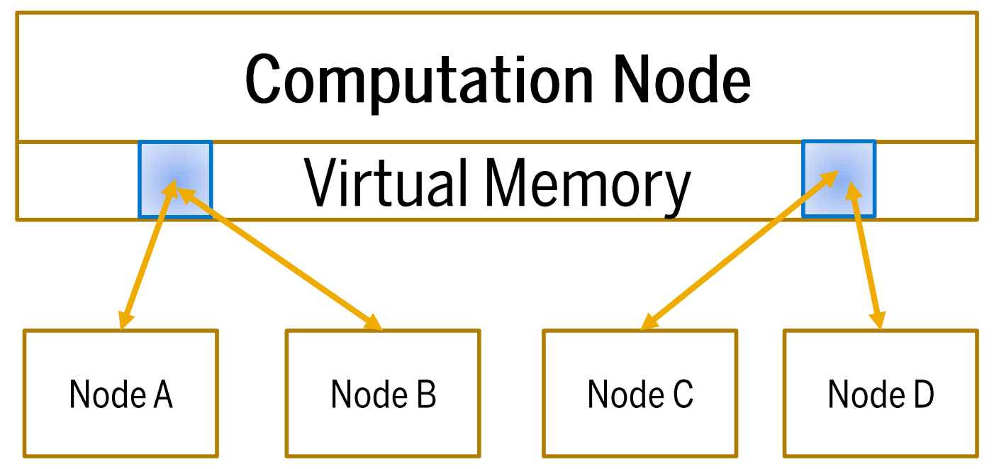
Performance: Minimal Overhead



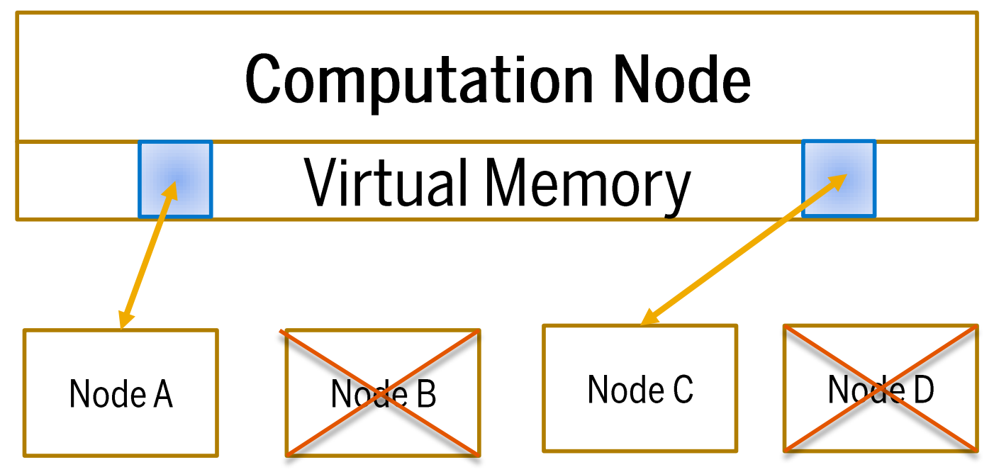
Performance evaluation was run on a compute node with 4GB RAM, and 2x remote nodes with 8GB RAM, communicating via SoftiWarp over TCP/IP.

Failover and Re-entrance sequence

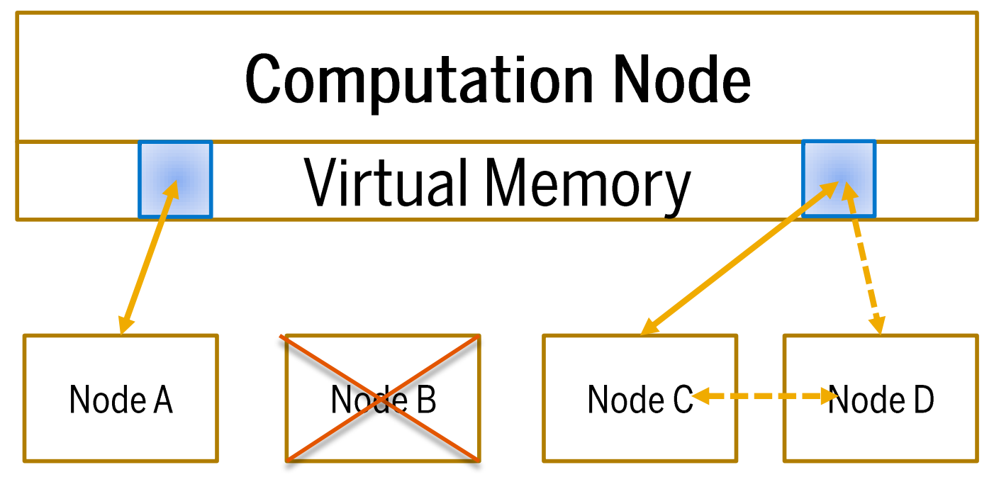
1. Every page backed by [2…n] remote nodes.



2. Failure in remote node has no immediate effect.



3. Linear sync process for remote node re-entrance.



Future Work

RRAIM is a part of the Hecatonchire Project – which goals are:

- Full resource liberation of data center
- Breaking down nodes into basic elements (CPU, Memory, I/O)
- Eliminate limitation of current cloud paradigm
- Seamlessly integrate with existing technology

We would like to thank John Stewart and David Leib for their contribution in the Hecatonchire project. We would also like to thank Dr. Eliezer Levy for his support in designing RRAIM.

References and Related Work

1) D. Magenheimer, RAMster: Peer-to-peer Transcendent Memory, http://marc.info/?l=linux-mm&m=130013567810410
2) M. R. Hines and K. Gopalan. 2007. MemX: supporting large memory workloads in Xen virtual machines. In *Proceedings of the 2nd international workshop on Virtualization technology in distributed computing* (VTDC '07)
3) D. G. Andersen, J. Franklin, M. Kaminsky, A. Phanishayee, L. Tan, and V. Vasudevan. 2009. FAWN: a fast array of wimpy nodes. In *Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles* (SOSP '09)
4) J. Ousterhout, P. Agrawal, D. Erickson, C. Kozyrakis, J. Leverich, D. Mazieres, S. Mitra, A. Narayanan, G. Parulkar, M. Rosenblum, S. M. Rumble, E. Stratmann, R. Stutsman. 2010. The case for RAMClouds: scalable high-performance storage entirely in DRAM. *SIGOPS Oper. Syst. Rev.* 43, 4.
5) Softiwarp: A software iWARP driver for OpenFabrics B. Metzler, F. Neeser, P. Frey, OpenFabrics Sonoma Conference, March 2009.
6) RDMA over Converged Ethernet, http://institute.lanl.gov/isti/summer-school/cluster network/projects-2010/Team CYAN Implementation and Comparison of RDMA Over Ethernet Presentation
7) SFW OFED Distribution with Soft-RoCE in Linux, http://www.systemfabricworks.com/news/system-fabric-works-ofeddistribution-supports-rdma-over-converged-ethernet-roce