Scale-Out NUMA

Stanko Novaković,[†] Alexandros Daglis,[†] Edouard Bugnion,[†] Babak Falsafi,[†] and Boris Grot[‡]

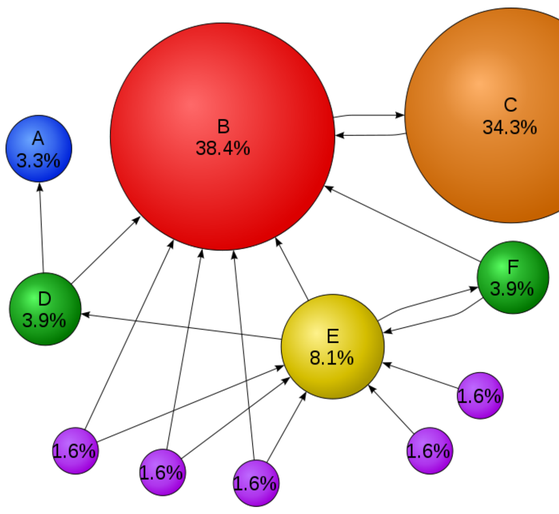
[†]EcoCloud, EPFL [‡]University of Edinburgh

1. Web-scale apps' characteristics

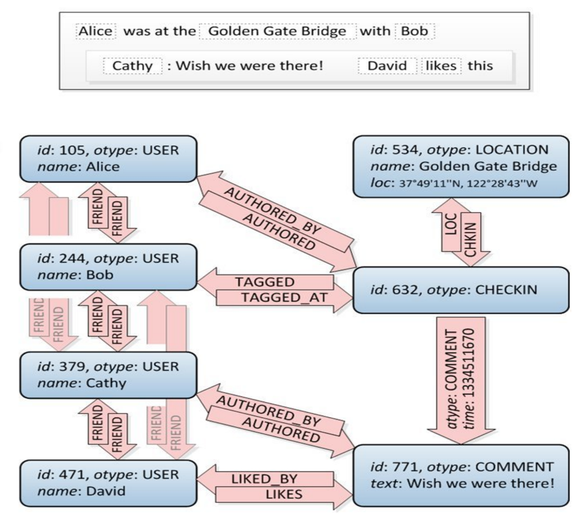
➢ **Offline/Online analytics and online query processing**

➢ **Common properties:**
- Large datasets, many nodes
- Frequent accesses to non-local data
- Individual nodes operate on small data items
- Little work per data item, measured in 100s of ns

Need communication latency ≈ computation latency



[Google PageRank]



[Facebook graph]

2. Existing approaches are ill-suited

➢ **Shared memory (ccNUMA)**
- \+ Low latency to remote data
- \- Limited scalability, high cost, single failure domain

➢ **Distributed memory using TCP/IP over Ethernet**
- \+ High scalability using commodity parts
- \- High remote access latencies (up to 1000x of local)

➢ **Distributed memory using RDMA over InfiniBand**
- \+ High scalability, low latency at datacenter-sustainable cost
- \- Latency of accessing remote memory still high (>10x of local)

3. Latency: The pain points

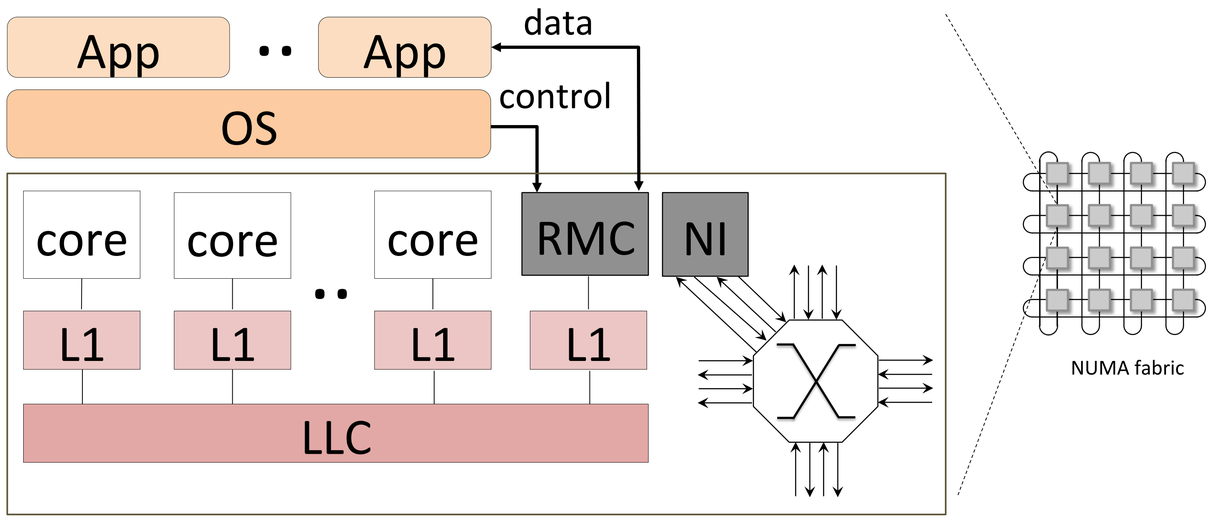
➢ **Deep Network Stacks**
- Significant computation per packet
- Major mismatch with emerging lean-core architectures

➢ **PCIe/DMA latencies limit performance**
- Up to 500ns overhead from transfers over PCIe alone
- Lack of coherence requires costly replication of data structures

4. Our proposal: Scale-Out NUMA

➢ **Rack-scale system architecture based on NUMA**
- Lean memory fabric
  - \+ Reliable interconnect with low-radix routers and wide links
- Integrated Remote Memory Controller (RMC)
  - \+ Cheap sharing of key in-memory data structures
- Global (partitioned) virtual address spaces
  - \+ Split-phase API for access to non-local data

➢ **Scale-out NUMA architecture**



➢ **How do we get to ultra-low latency?**
- Integration/local coherence for fast access to key data structures
- Kernel bypass allows user to directly interact with the RMC
- Low router pin-to-pin delays and fixed topology
- Small form factor with low inter-node distances

5. Evaluation

A) Emulation platform (for development and testing)
- ➢ Based on the Xen hypervisor and a ccNUMA server

B) Simulated hardware
- ➢ Flexus full-system, cycle-accurate simulation

Transport	soNUMA		RDMA/IB*
	Emulation	Simulation	
Max BW (Gbps)	1.8	77	50
Read RTT (μs)	1.5	**0.3**	1.19
Fetch-and-add (μs)	1.5	0.3	1.15
IOPS (Mops/s)	1.97	10.9	35 @ 4 cores

*Mellanox ConnectX-3 RDMA host channel adapter on host Xeon E5-2670 2.60Ghz via PCIe-Gen3

Remote latency at small factor of local memory latency

➢ **Graph processing (PageRank) study**