nano-tera.ch
swiss scientific initiative in health / security / environment systems

YINS RTD 2013

FNSNF

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

PARSA PARALLEL SYSTEMS ARCHITECTURE LAB

DCSL

Scale-Out NUMA

Stanko Novakovic, Alexandros Daglis, Boris Grot, Edouard Bugnion, Babak Falsafi*

EPFL, University of Edinburgh*

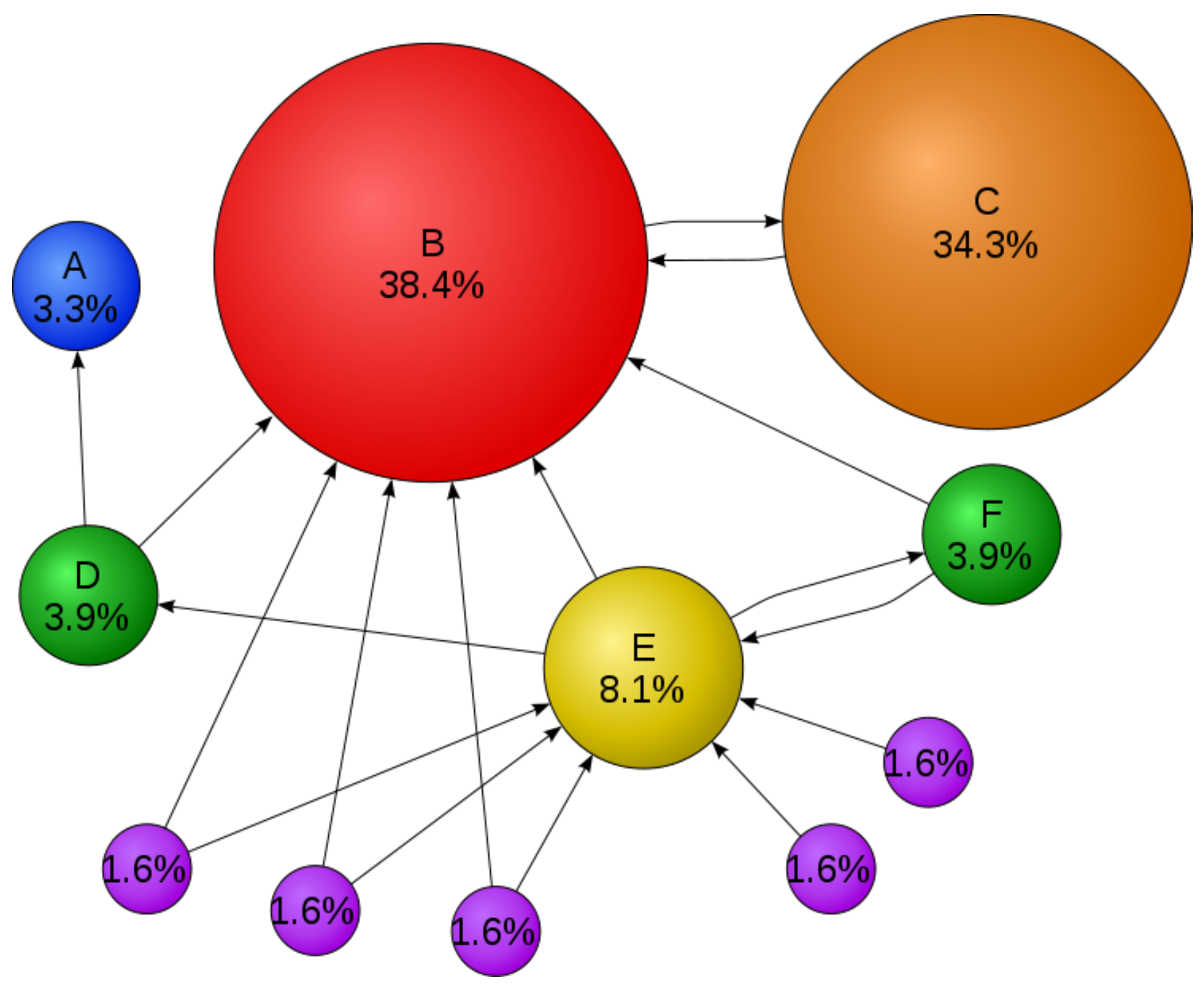
1. Large-scale datacenter applications

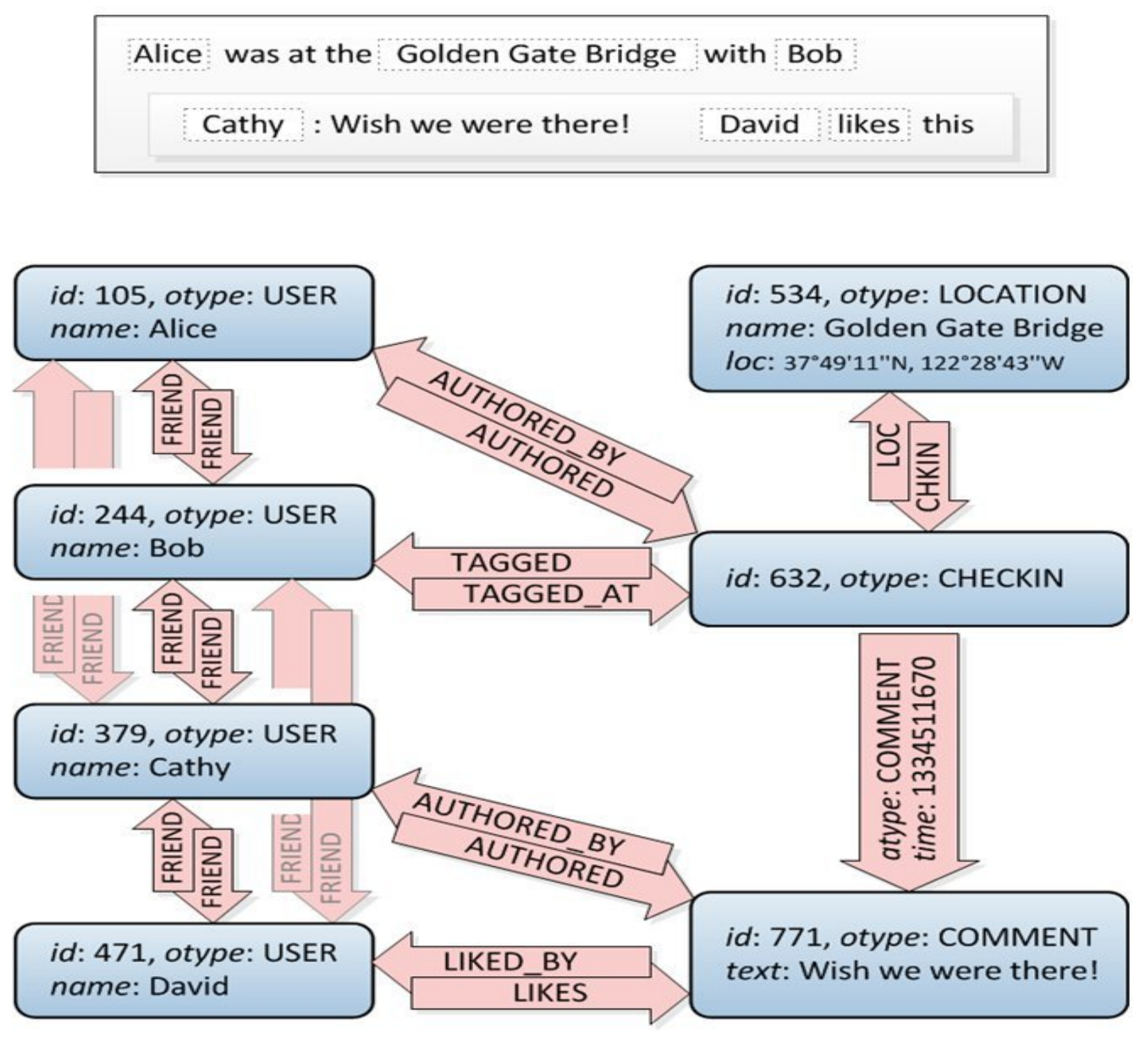
Big-data analytics and data serving

Common properties:

- Large datasets, many nodes
- Frequent accesses to non-local data
- Very little processing per query/algorithm iteration

Most DC applications are network-bound




Alice was at the Golden Gate Bridge with Bob
Cathy : Wish we were there! David likes this
id: 105, otype: USER name: Alice
id: 534, otype: LOCATION name: Golden Gate Bridge loc: 37°49'11"N, 122°28'43"W
id: 244, otype: USER name: Bob
id: 632, otype: CHECKIN
id: 379, otype: USER name: Cathy
id: 471, otype: USER name: David
id: 771, otype: COMMENT text: Wish we were there!
FRIEND
FRIEND
AUTHORED_BY
AUTHORED
TAGGED
TAGGED_AT
LOC
CHKIN
atype: COMMENT time: 1334511670
LIKED_BY
LIKES

2. Existing approaches 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

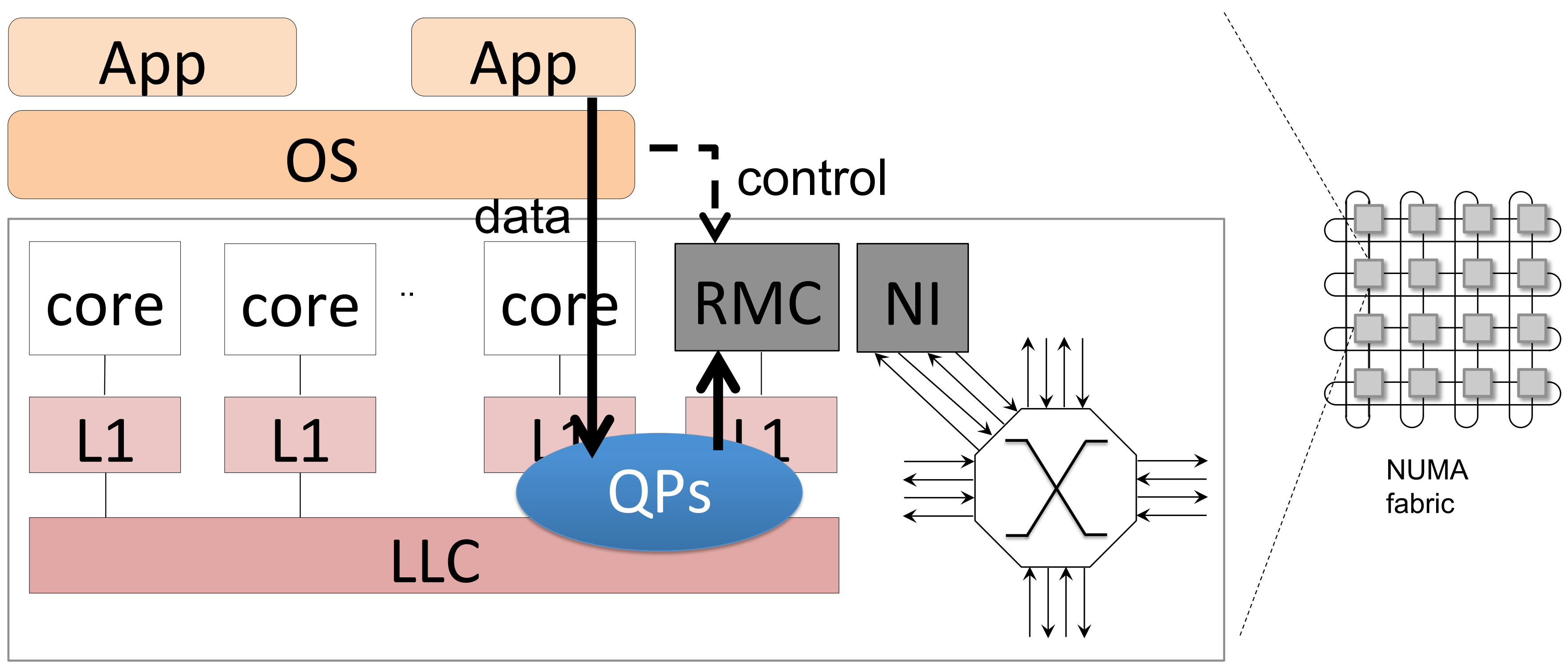
+ High scalability, low latency
- Remote access latency memory still high (>10x of local)

Deep network stacks, PCI/DMA limit performance

3. Our proposal: Scale-Out NUMA

Rack-scale system based on NUMA transport

- Reliable and wide interconnect
- Integrated (locally) cache-coherent Remote MC
- Direct access via memory-mapped queue-pairs (QP)



Remote access latency of 300ns, DDR rate, scalable

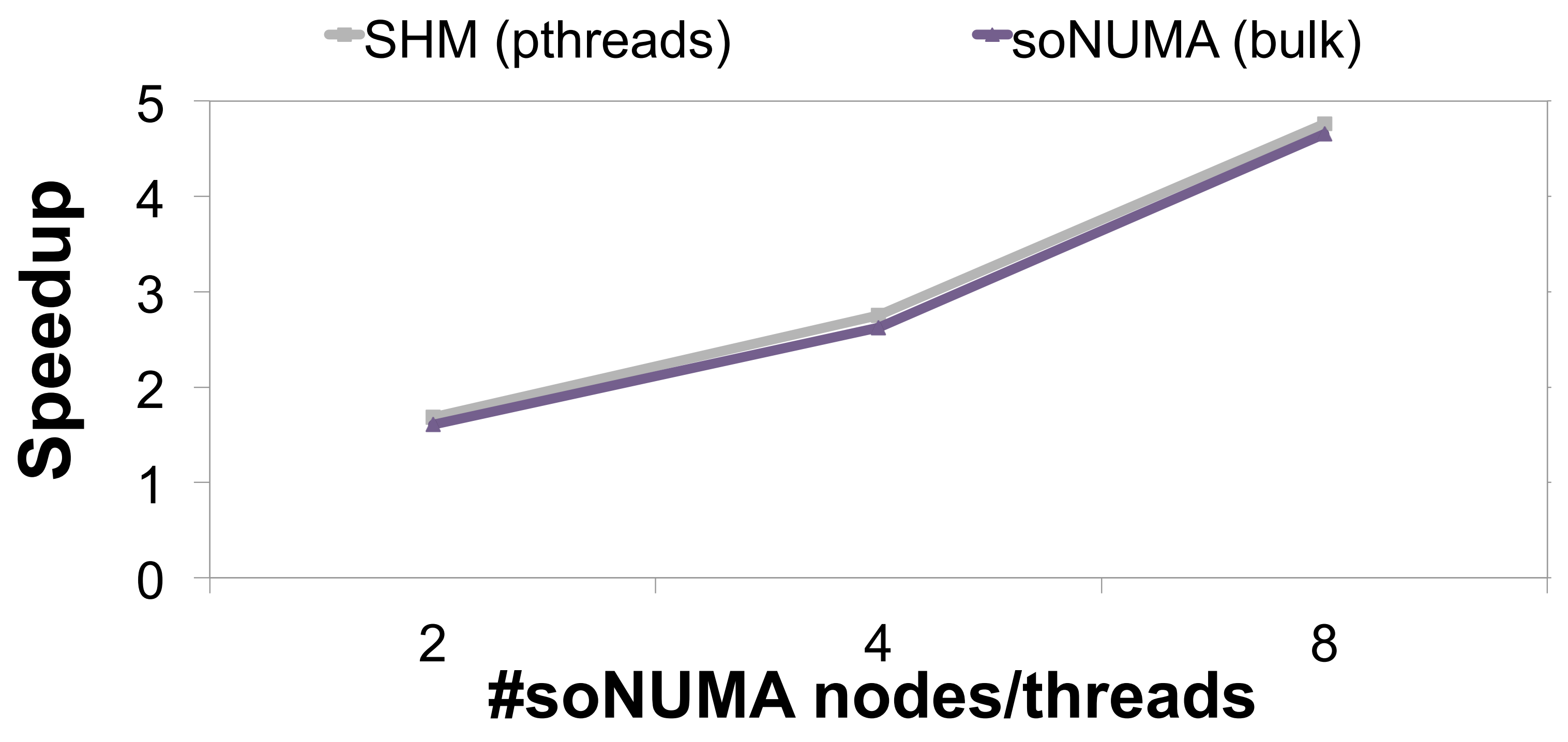
4. Rack-scale graph processing

Bulk Synchronous Parallel processing

- Iterative computation
- Servers exchange graph updates across iterations

PageRank on Twitter graph study

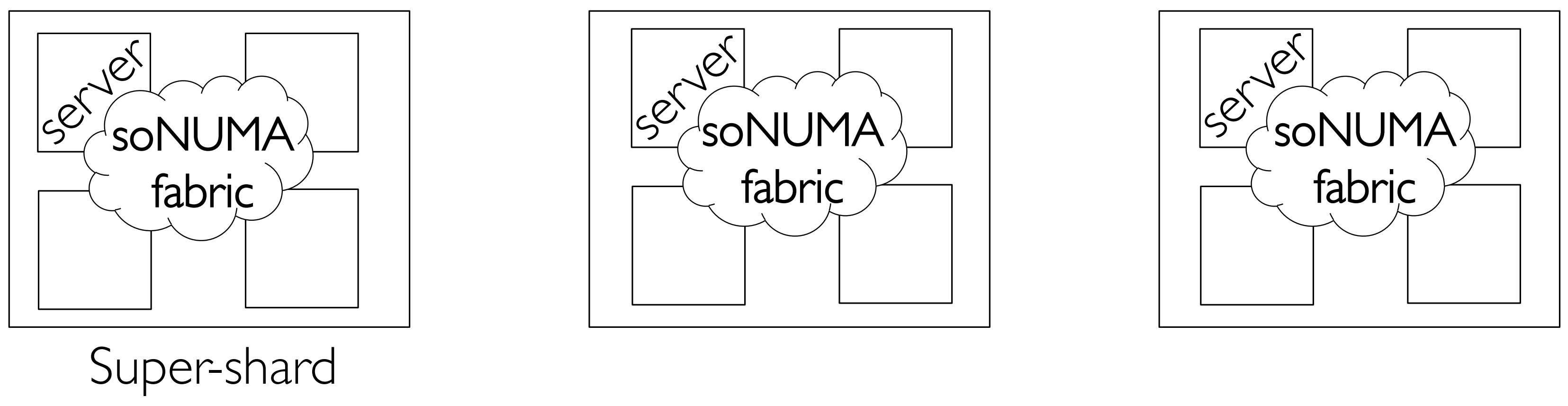
- soNUMA (bulk) → BSP implementation on soNUMA
- SHM (pthreads) → shared-memory implementation

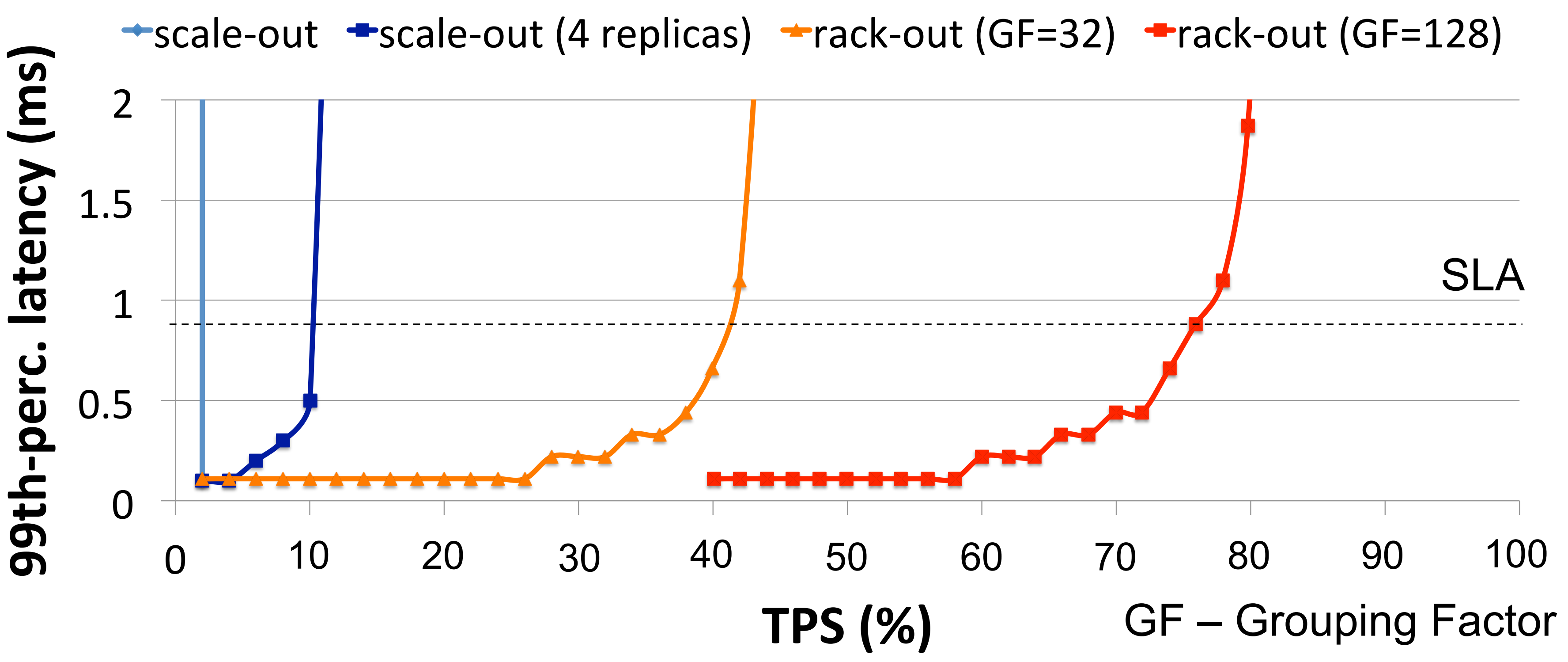


5. Rack-out data serving

Rack-out: shard data at rack-scale granularity

- Skewed access patters create hotspots in scale-out
- Group servers into racks → more compute/network
  → Deliver higher throughput w/o violating SLA


server
soNUMA fabric
Super-shard
server
soNUMA fabric
server
soNUMA fabric


scale-out
scale-out (4 replicas)
rack-out (GF=32)
rack-out (GF=128)
99th-perc. latency (ms)
SLA
TPS (%)
GF – Grouping Factor