Georgia Tech Computer Science

Optimizing the "Last Mile" with Network-Compute Co-Design

Alexandros (Alex) Daglis
Assistant Professor, School of Computer Science
cc.gatech.edu/~adaglis

CNRCH Summit, January 29th 2021

Large-Scale Online Services

Online services live in massive datacenters

- 10,000s of servers

Tight quality guarantees (SLOs)

- Care about "worst-case" (tail) latency

Data distributed across thousands of servers

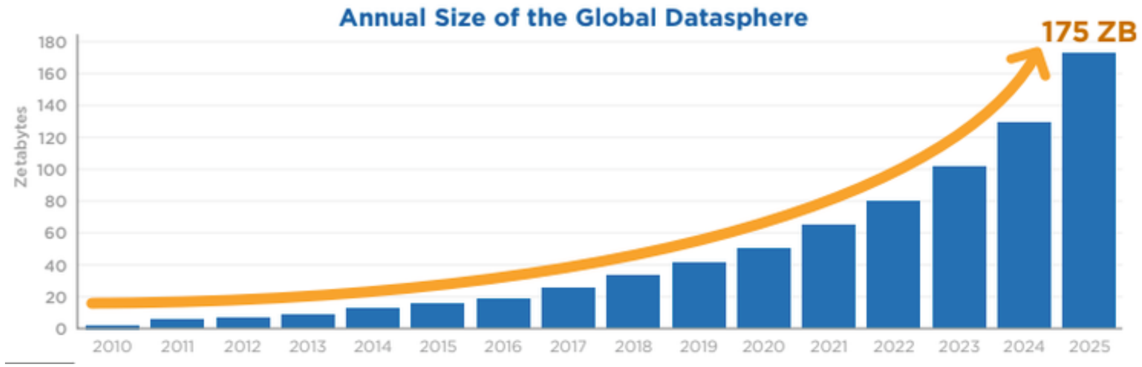
Servers communicate over datacenter network

Distribution enables scalable performance & low response latency

Growing Pressure on the Network

Trend I: More data → more scale-out

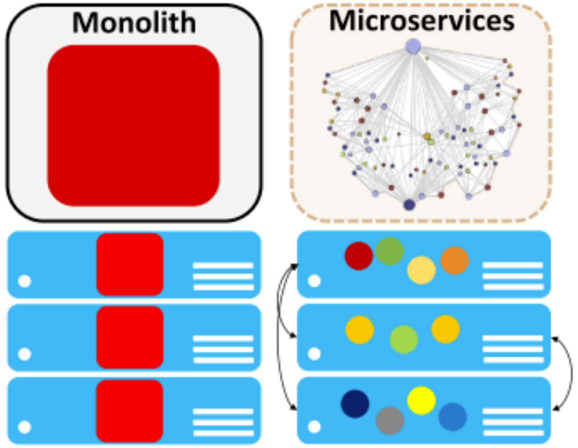
- 66% growth per year



Credits: IDC, 2018

Trend II: Software decomposition (microservices, serverless)

- Service times in µs domain
  → network message every few k CPU cycles!



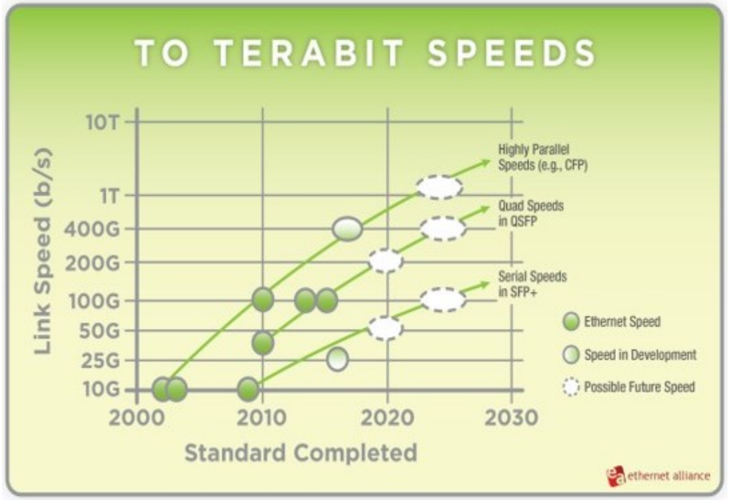
Credits: Gan, ASPLOS'19

Network emerging as key performance determinant

Datacenters Keeping up with Growing Demand

Growing bandwidth & high path diversity

- Datacenter-wide roundtrips <20µs



Credits: Ethernet Alliance, 2015

Optimized protocols cut messaging costs

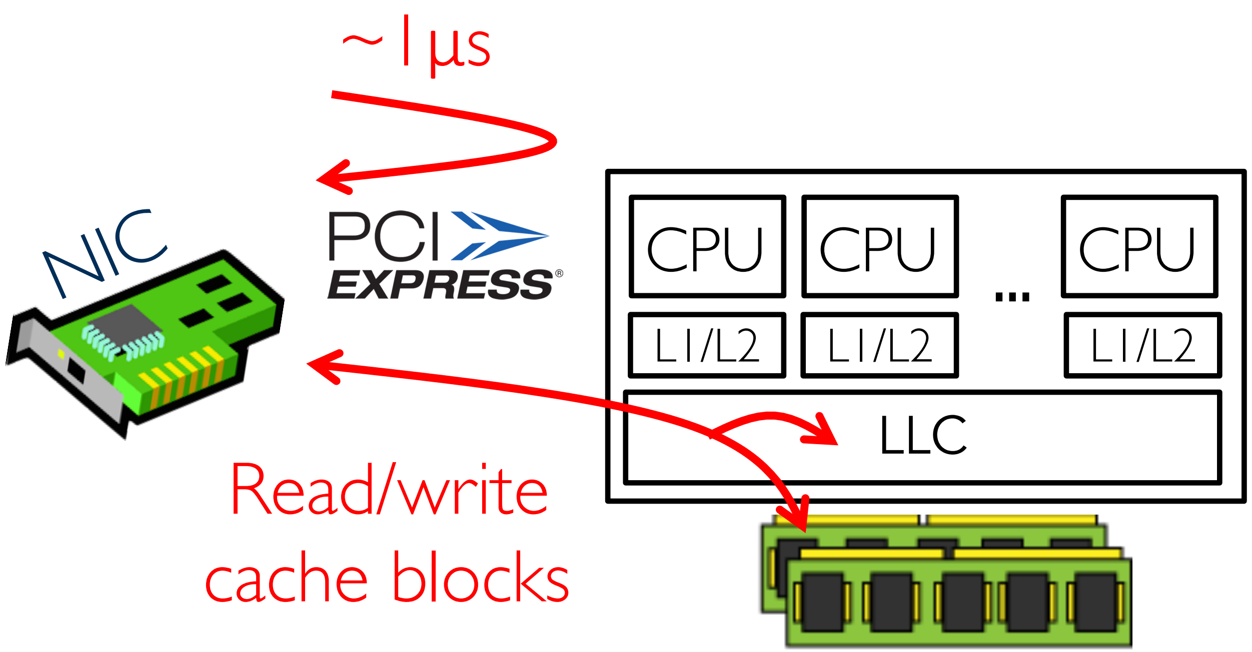
- From 10+µs to sub-µs

...despite networking evolution, NIC-CPU interface still architected as IO

- Bandwidth-optimized, high latency
- Performance and semantic obstacles

Need architectural revisiting of "last mile"

What's Wrong in this "Last Mile"?



Need

- New interfaces
- Richer operations
- Advanced interactions with compute & memory

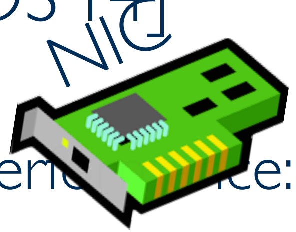
Making the NIC a First-Class Citizen

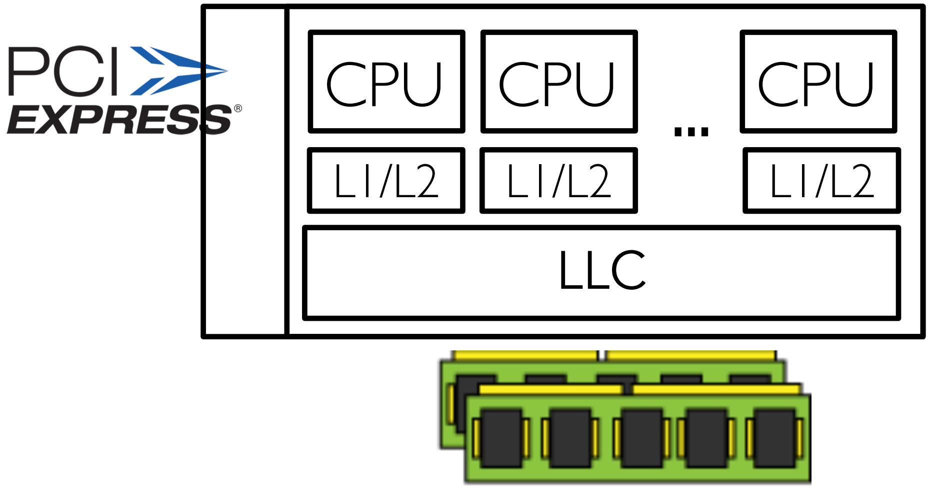
New Interfaces
Richer Operations
Advanced Interactions

Optimize Network-Compute interface via NIC integration

Scale-Out NUMA architecture [ASPLOS'14]

- NIC in coherence domain
- Rack-scale scale-out systems w/ NUMA performance: remote memory within ~3x of local





More than immediate latency gains

- Paves way for higher-level operations with richer semantics

Integration facilitates network-compute co-design

From Cache Blocks to Memory Objects

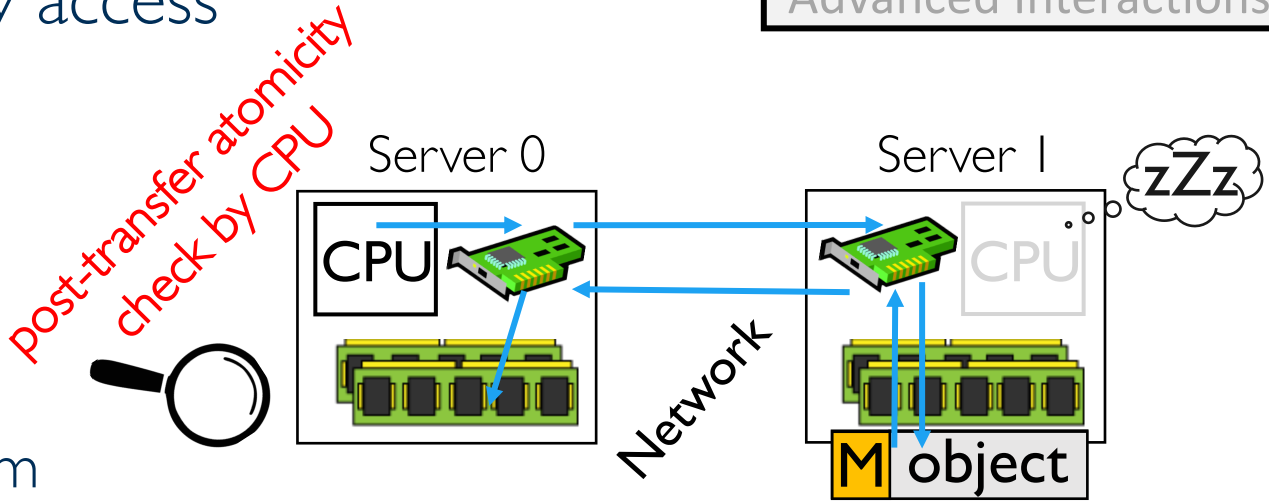
New Interfaces
Richer Operations
Advanced Interactions

RDMA enables direct remote memory access

- Great for distributed object stores

No object-level atomicity guarantees with basic "read" primitive!

- Need out-of-band verification mechanism



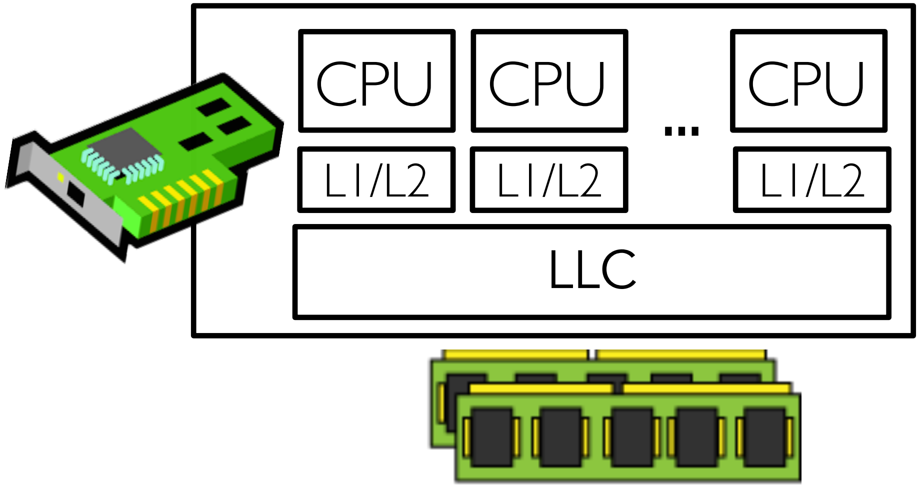
NIC in coherence domain → snoop coherence traffic to target object [MICRO'16]

- On-the-fly atomicity check, no software involvement
- 35-50% faster atomic object reads from remote memory

Tight NIC-compute coupling enables Atomic Object Read hardware primitive

NIC-driven Load Balancing Opportunities

Packet distribution to cores critical for scalability

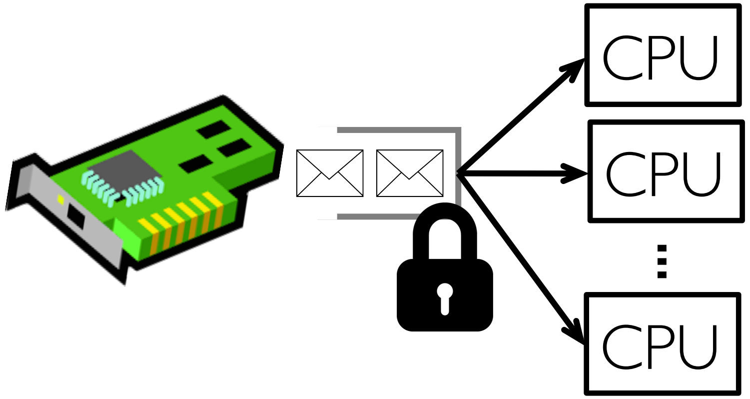


NIC-driven Load Balancing Opportunities

New Interfaces
Richer Operations
Advanced Interactions

Packet distribution to cores critical for scalability

- Software-based mechanisms expensive for µs-scale services



NIC-driven Load Balancing Opportunities

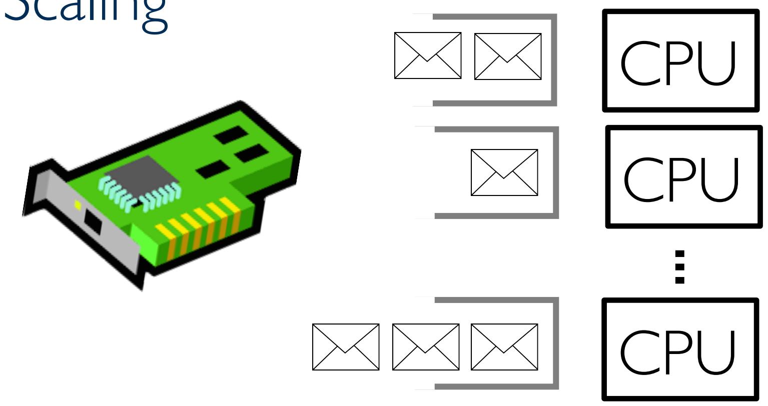
New Interfaces
Richer Operations
Advanced Interactions

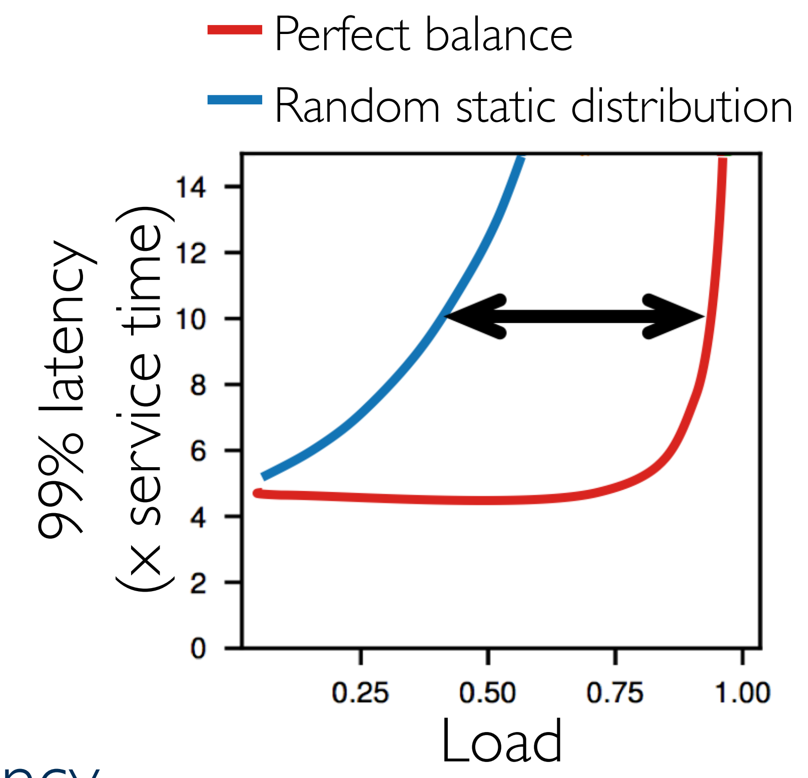
Packet distribution to cores critical for scalability

- Software-based mechanisms expensive for µs-scale services

NIC can spread incoming load to cores

- e.g., Receive-Side Scaling





But static decisions → load imbalance → hurts tail latency

Integration Facilitates Load Balancing

New Interfaces

Richer Operations

Advanced Interactions

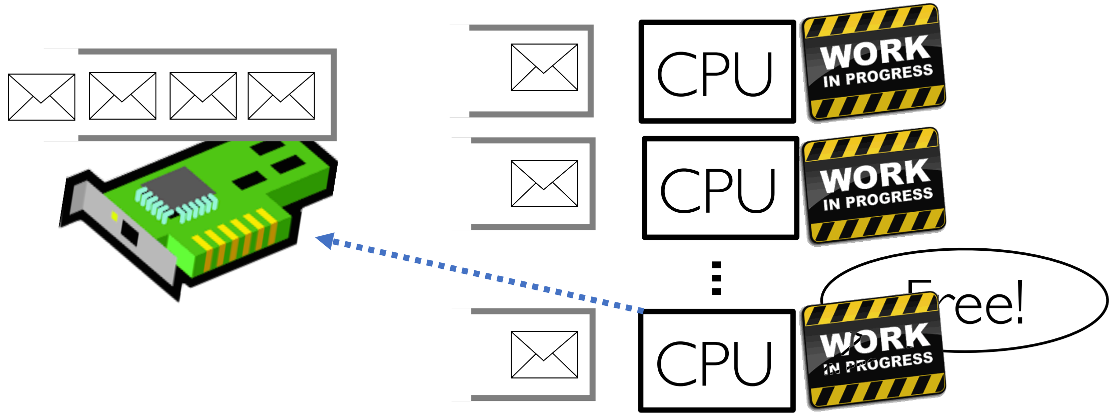
Co-design NIC with compute

- Direct interaction and load monitoring – dispatch work when compute available

[ASPLOS'19]

Simple greedy approach works even for µs-scale services due to integration

- Nanosecond-scale on-chip latencies

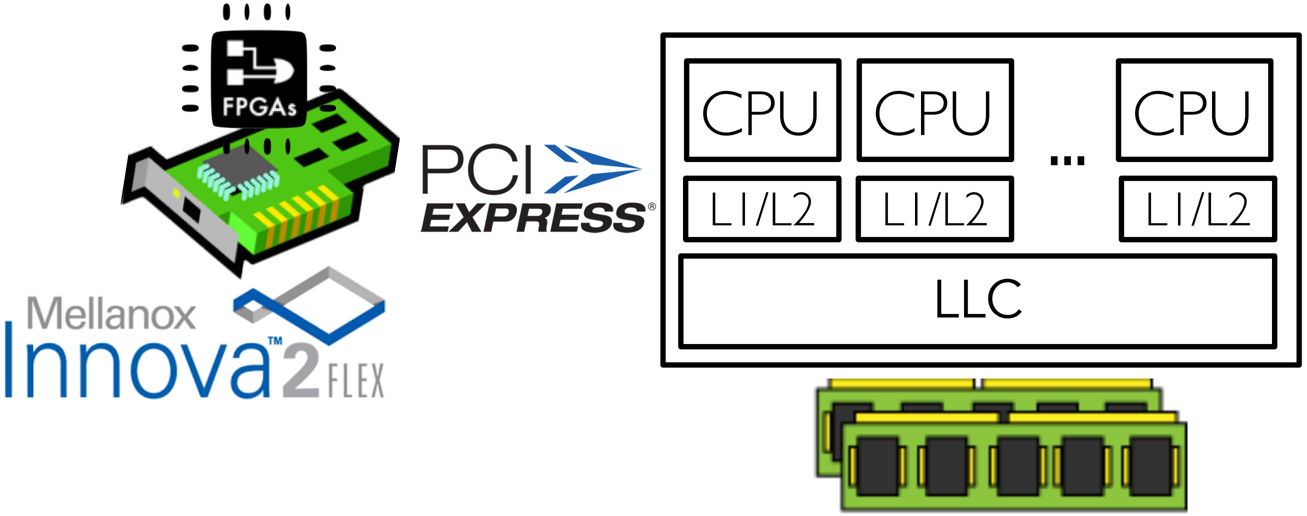


20-40% higher throughput under SLO for µs-scale services

NIC-driven Load Balancing Extensions

Is NIC-driven load balancing applicable to existing smartNICs?

- PCIe latency precludes greedy approach
- But can learn and dynamically approximate per-core load



Decisions under workload diversity – can NIC predict service time?

Advanced Interactions via Co-Design

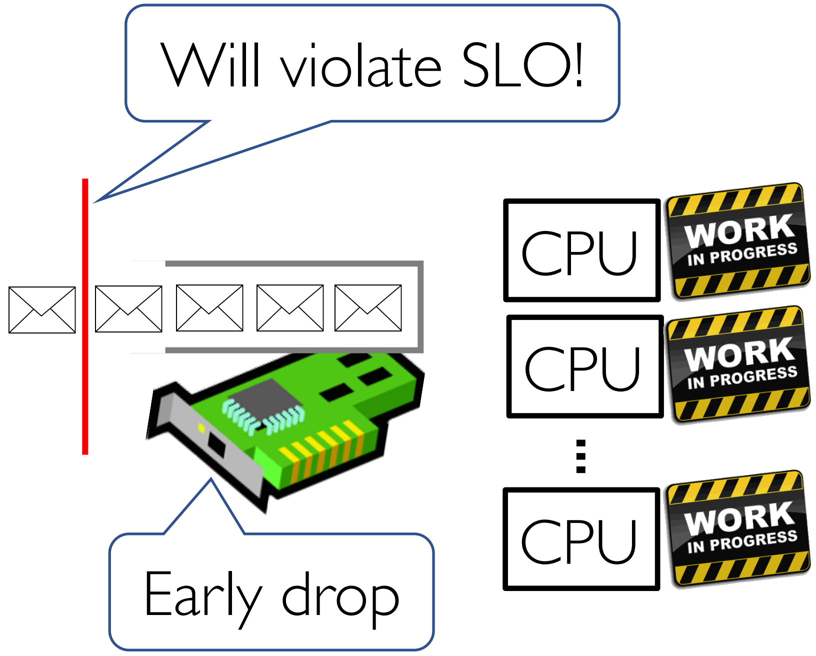
New Interfaces
Richer Operations
Advanced Interactions

Software hints enable data movement optimizations

Expose application service times and SLO to NIC

Enable SLO-aware packet management [ISCA'20]

- Minimize data movement under high contention
- Prevent spill of latency-sensitive traffic to DRAM



Up to 2x throughput with SLO-aware packet management

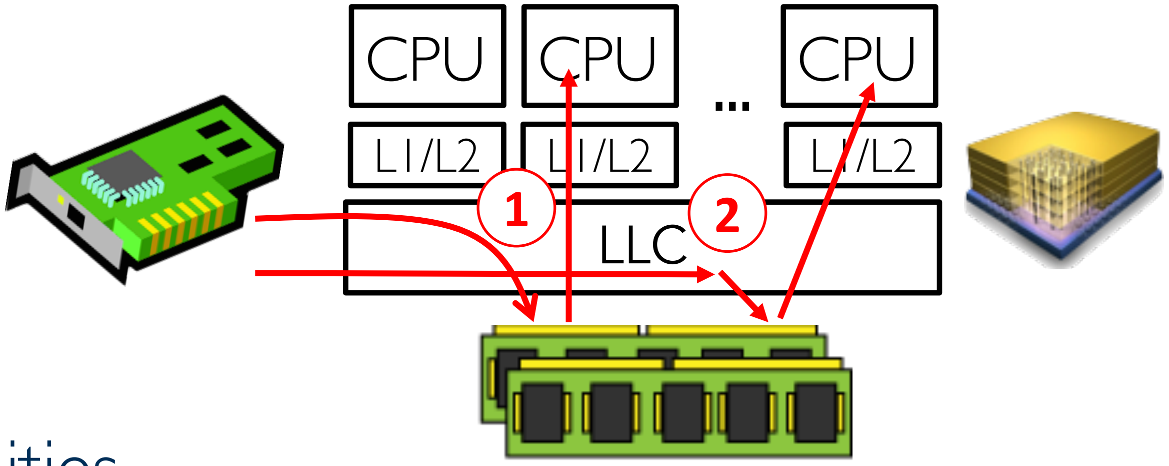
Judicious Data Movement

New Interfaces
Richer Operations
Advanced Interactions

Incoming data placement policies are static – and suboptimal

Mainstream approaches:

1. Data into DRAM
2. Data into subset of LLC (DDIO)



Data movement optimization opportunities

- Application-driven dynamic placement decision: DRAM, LLC, or private upper-level caches
- Even more interesting in heterogeneous, accelerator-rich architectures [ISCA'20, CAL'20]
- Header/payload splitting and separate manipulation

Ample opportunities in smarter data steering decisions

Conclusion

Evolution of online services puts network communication in spotlight

Advancements in networking technologies and protocols aligned with needs

… but also need architectural rethink for the "last mile"

Optimize network-compute-memory interactions via co-design

New Interfaces
Richer Operations
Advanced Interactions