Challenges and Opportunities in Exascale-Computing Interconnects

Manolis Katevenis and Nikolaos Chrysos

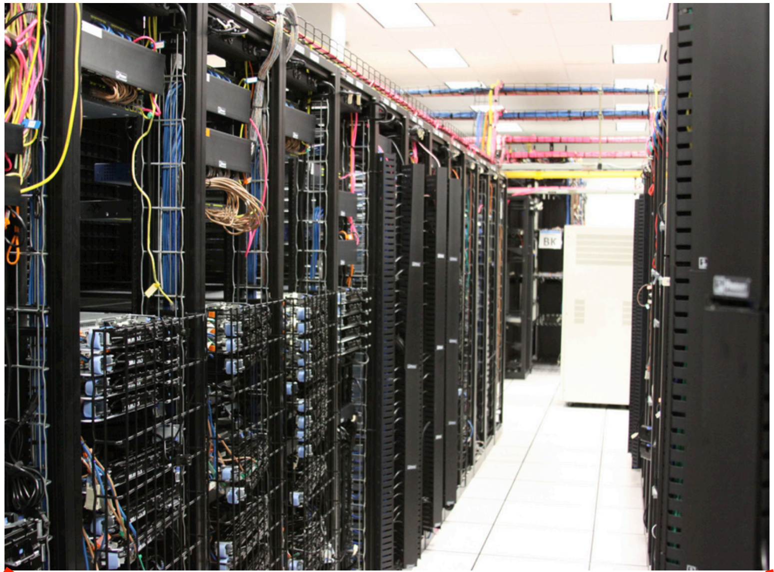
FORTH-ICS and Univ. of Crete, Greece

1st AISTECS Workshop, 18 January 2016, Prague, Czech Republic

Outline

- Warehouse-scale datacenters and supercomputers
- Traffic characteristics in commercial datacenters
- Efficient congestion control: an old, unresolved problem
- Multipathing: benefits & issues
- RDMA: optimizing data copying
- Global virtual address space & routing

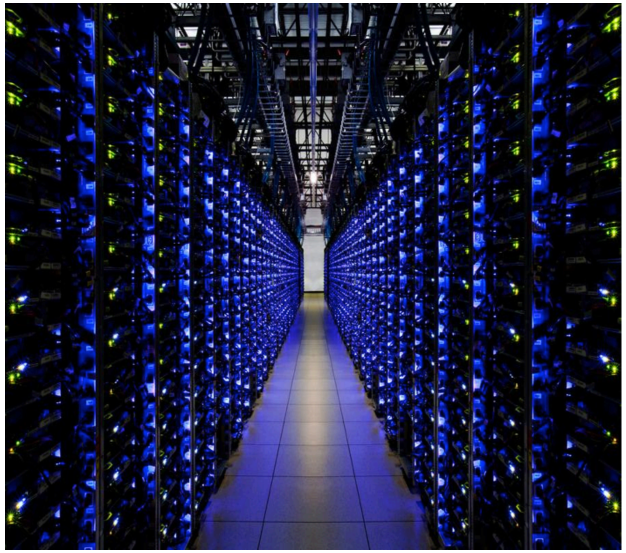
Datacenters: big-data stores of information society



What is a "datacenter"

A rack: 25-40 servers

Enterprise: a-few-hundred servers



Commercial datacenter:

Many-thousand servers

Supercomputers have similarities with datacenters

- Large scale installations
- Exploit massive parallelism
- … but designed to perfection
  - Custom one-off hardware vs. the economies of scale that rule in datacenters

Supercomputers have similarities with datacenters

- Large scale installations
- Exploit massive parallelism
- … but designed to perfection
  - Custom one-off hardware vs. the economies of scale that rule in datacenters

TOP500 list as of Nov. 2015

1. Tianhe-2 (MilkyWay-2) : 54.9 PFLOP/S @ 17.8 MW
2. Titan - Cray XK7 : 27.1 PFLOP/S @ 8.2 MW
3. Sequoia - BlueGene/Q : 20.1 PFLOP/S @ 7.9 MW

Supercomputers have similarities with datacenters

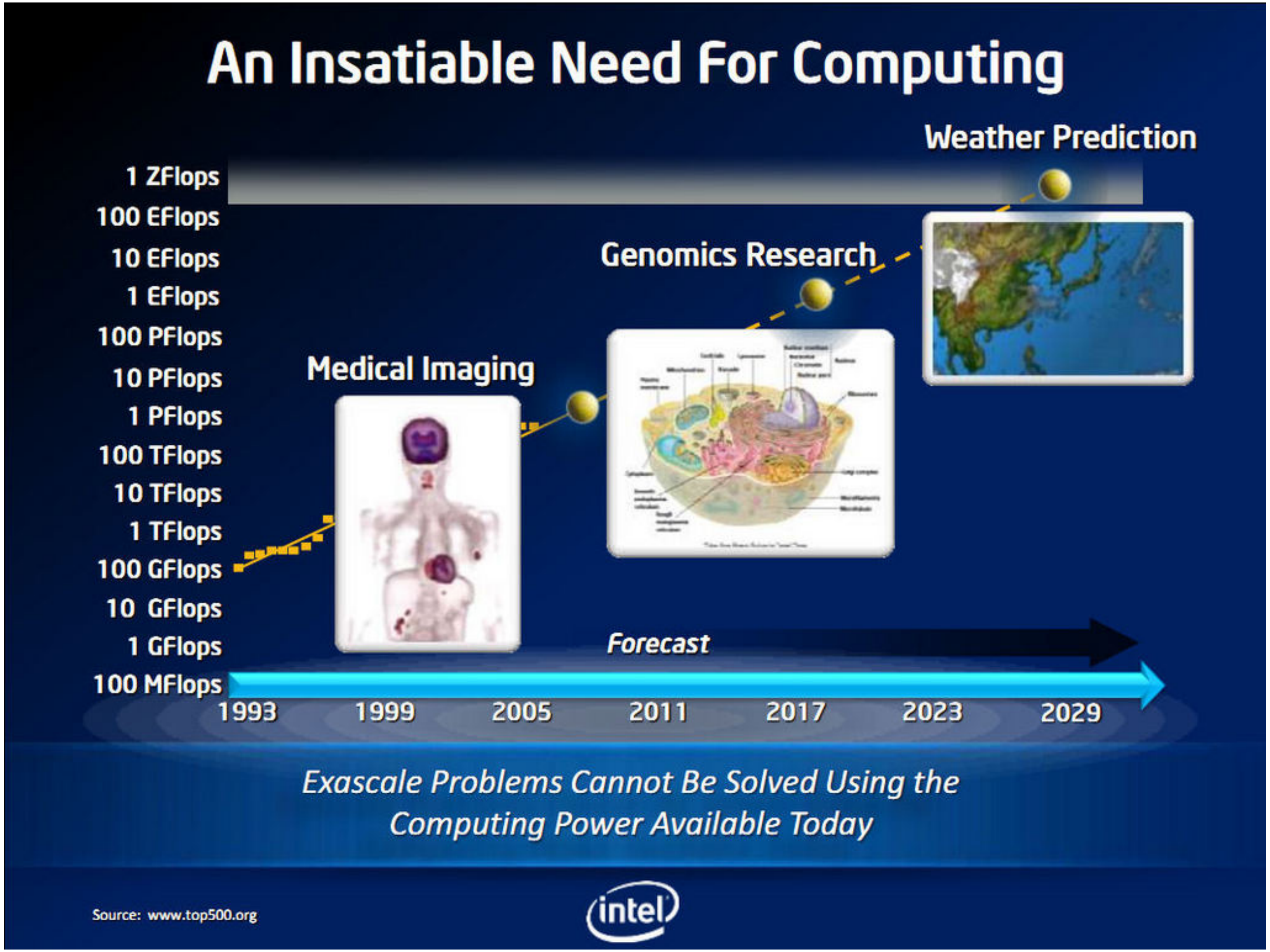
- Large scale installations
- Exploit massive parallelism
- … but designed to perfection
  - Custom one-off hardware vs. the economies of scale that rule in datacenters

TOP500 list as of Nov. 2015

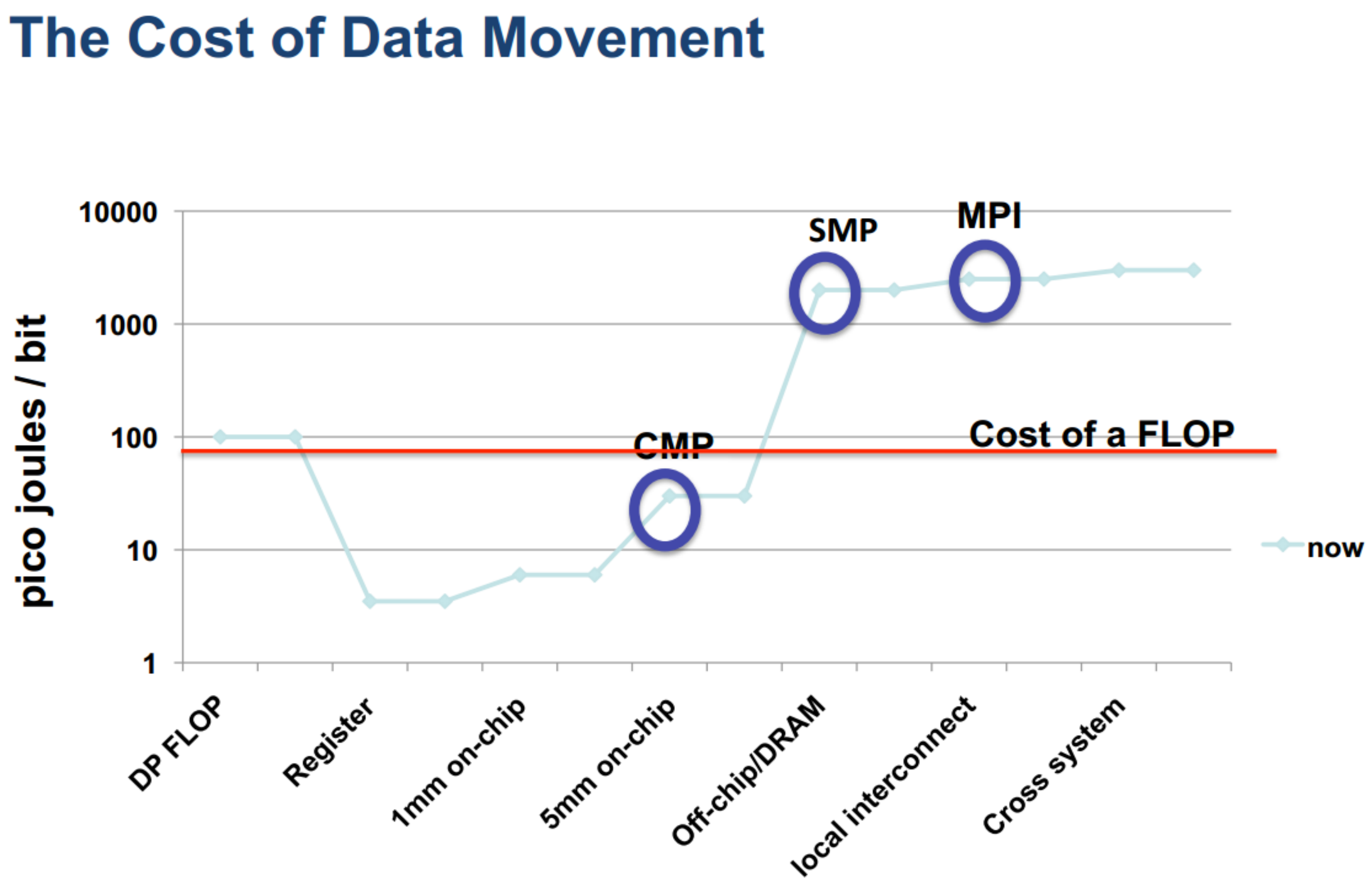
1. Tianhe-2 (MilkyWay-2) : 54.9 PFLOP/S @ 17.8 MW
2. Titan - Cray XK7 : 27.1 PFLOP/S @ 8.2 MW
3. Sequoia - BlueGene/Q : 20.1 PFLOP/S @ 7.9 MW

Google's power estimated at 40 PFLOPS in 8 Datacenters (2012)

Computing power must scale together w. society needs



Data movement is a massive wall in the road to exascale

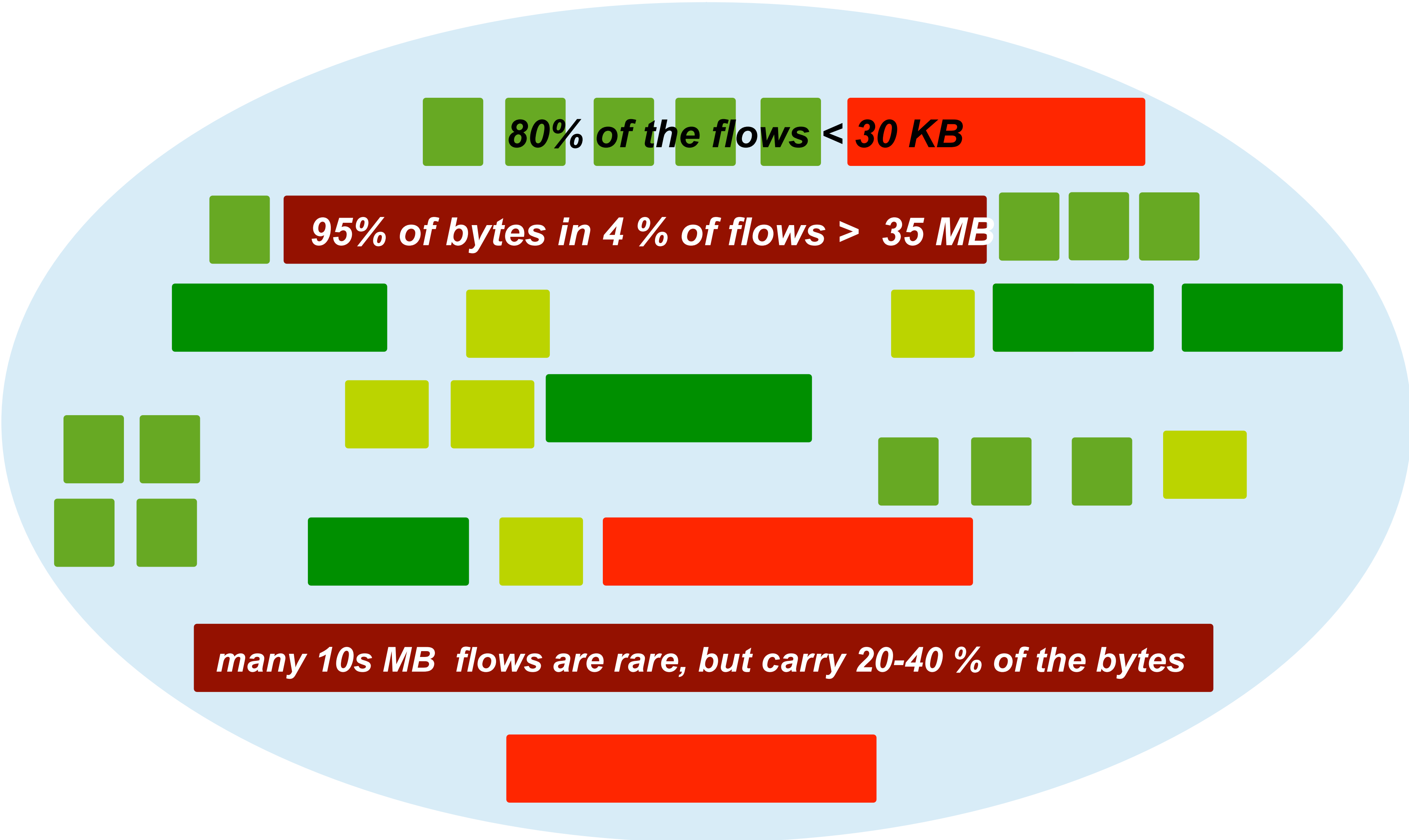


Courtesy: Horst Simon, Lawrence Berkeley NationalLaboratory

Outline

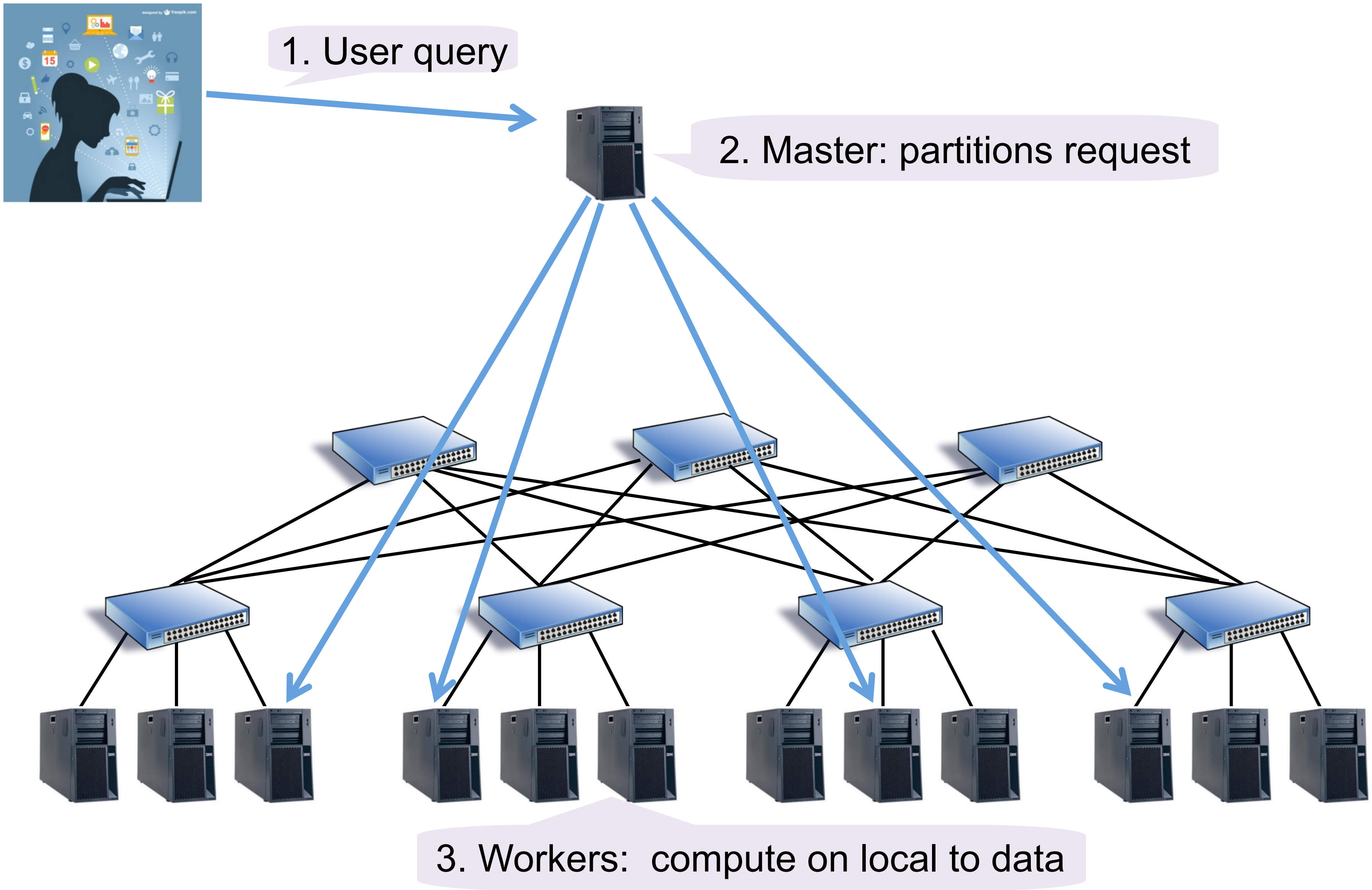
- Warehouse-scale datacenters and supercomputers
- **Traffic characteristics in commercial datacenters**
- Efficient congestion control: an old, unresolved problem
- Multipathing: benefits & issues
- RDMA: optimizing data copying
- Global virtual address space & routing

Flow sizes in commercial datacenters

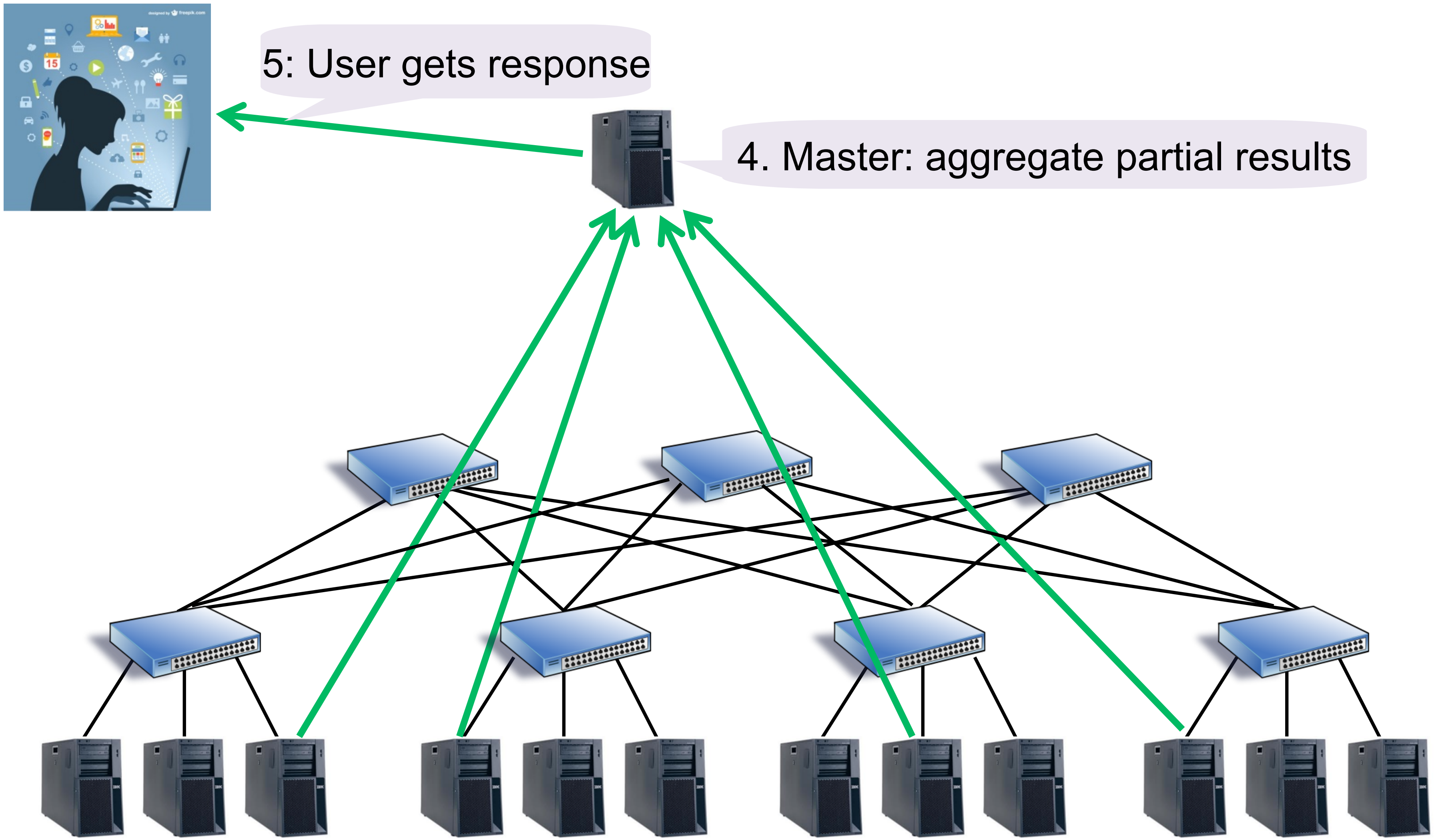


Sources of large flows: storage/ VM migration/ checkpointing..

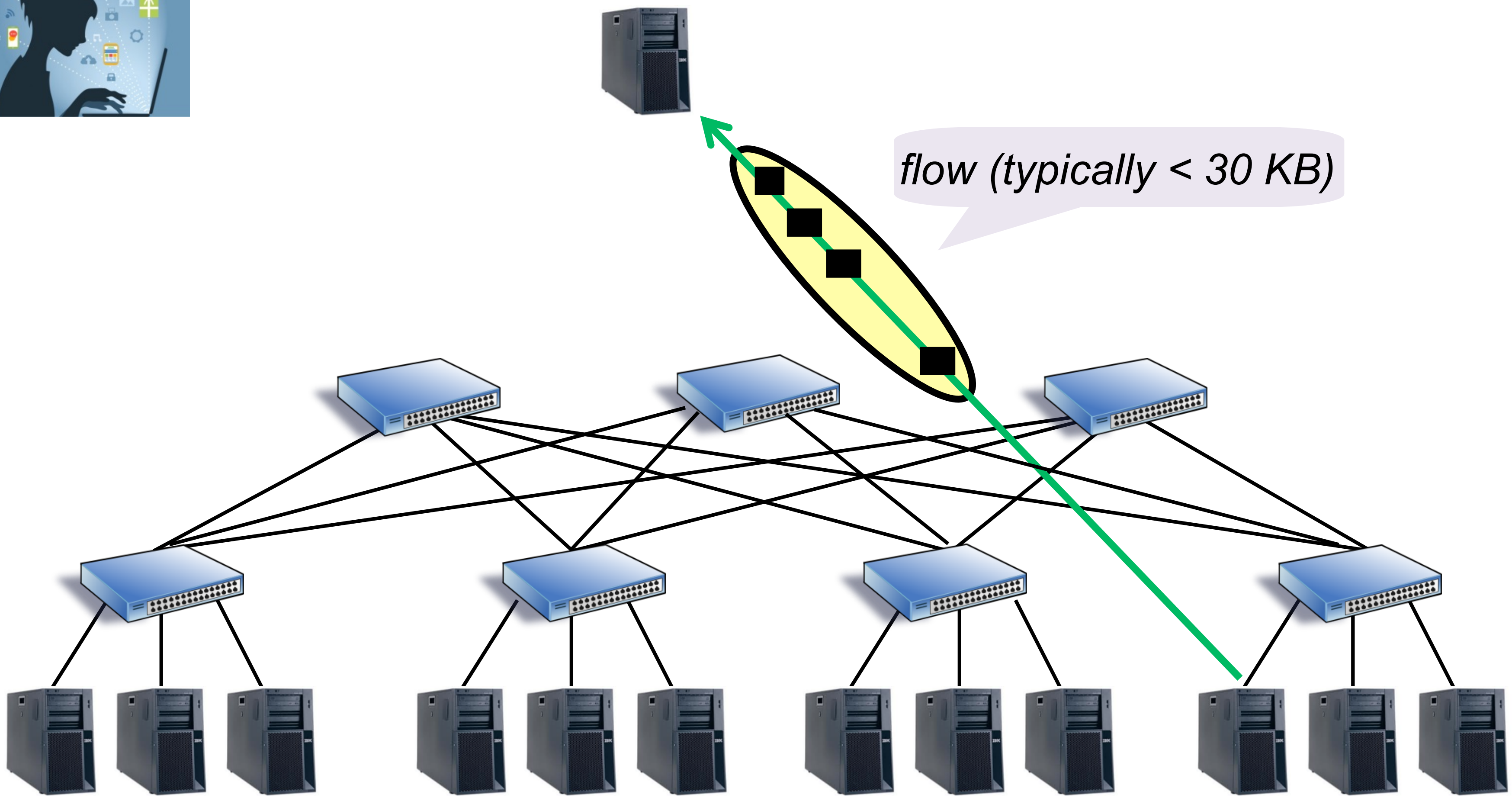
Online data intensive (OLDI) app's working on big-data



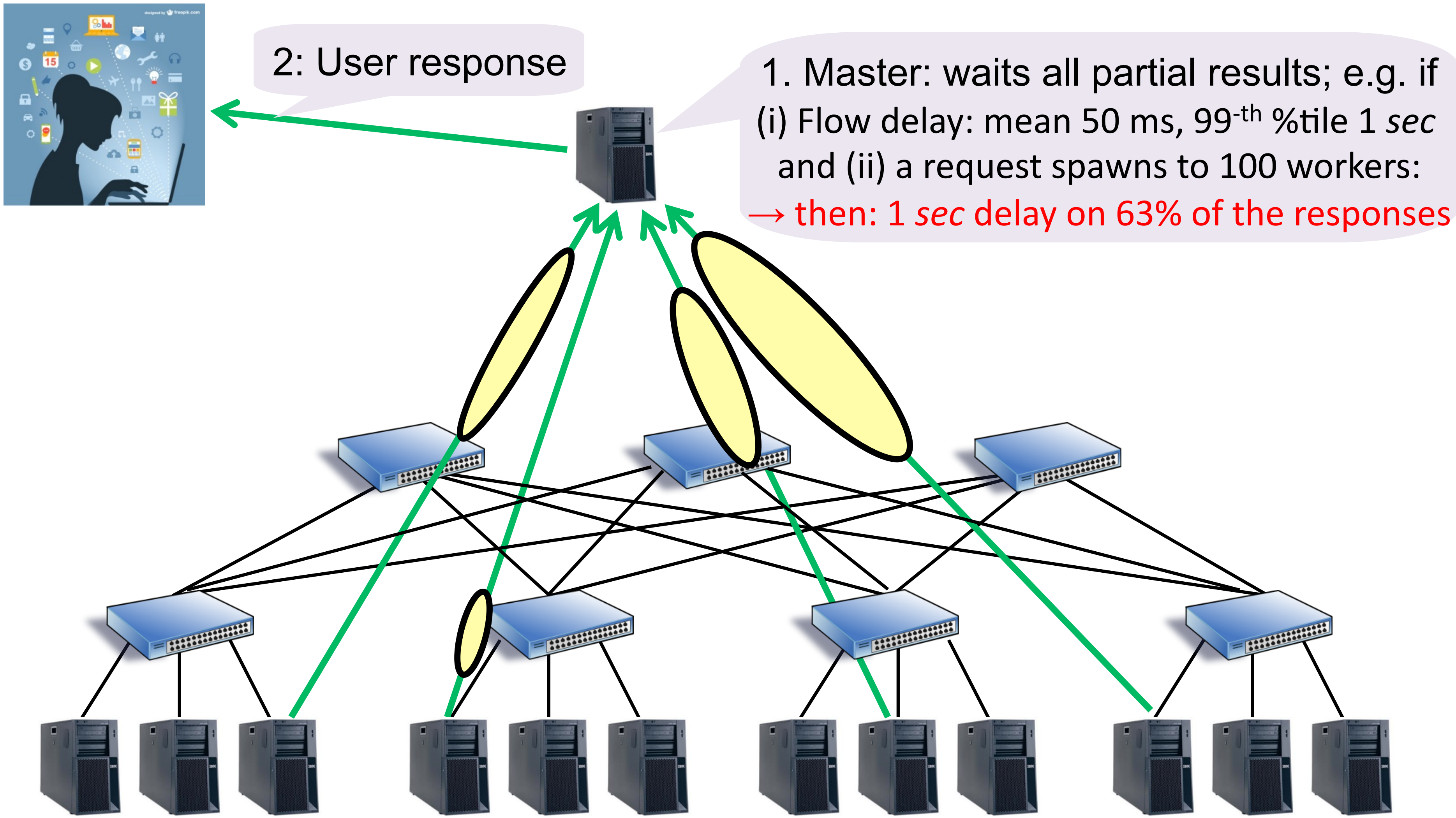
Online data intensive (OLDI) app's working on big-data



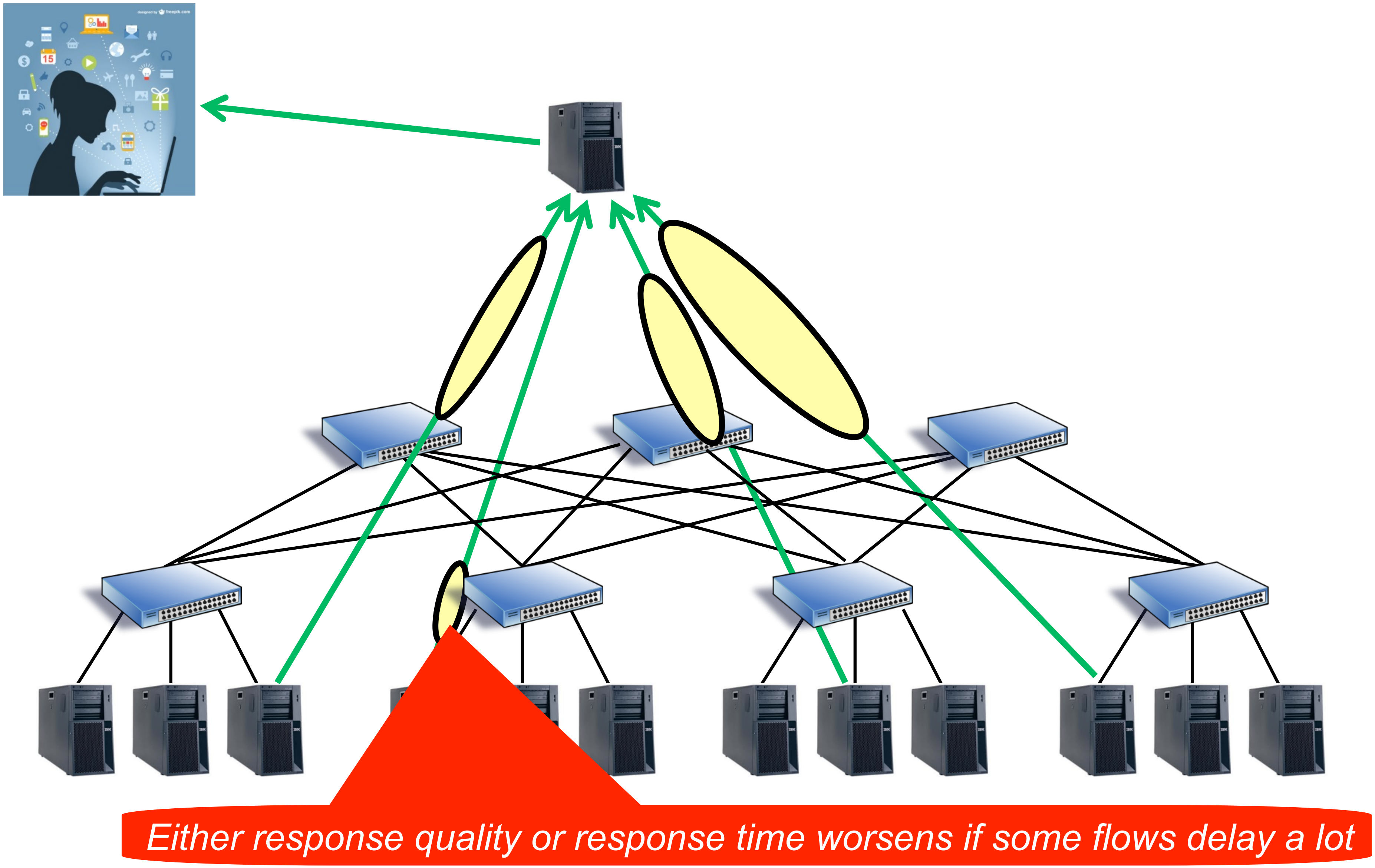
Partial results are usually small



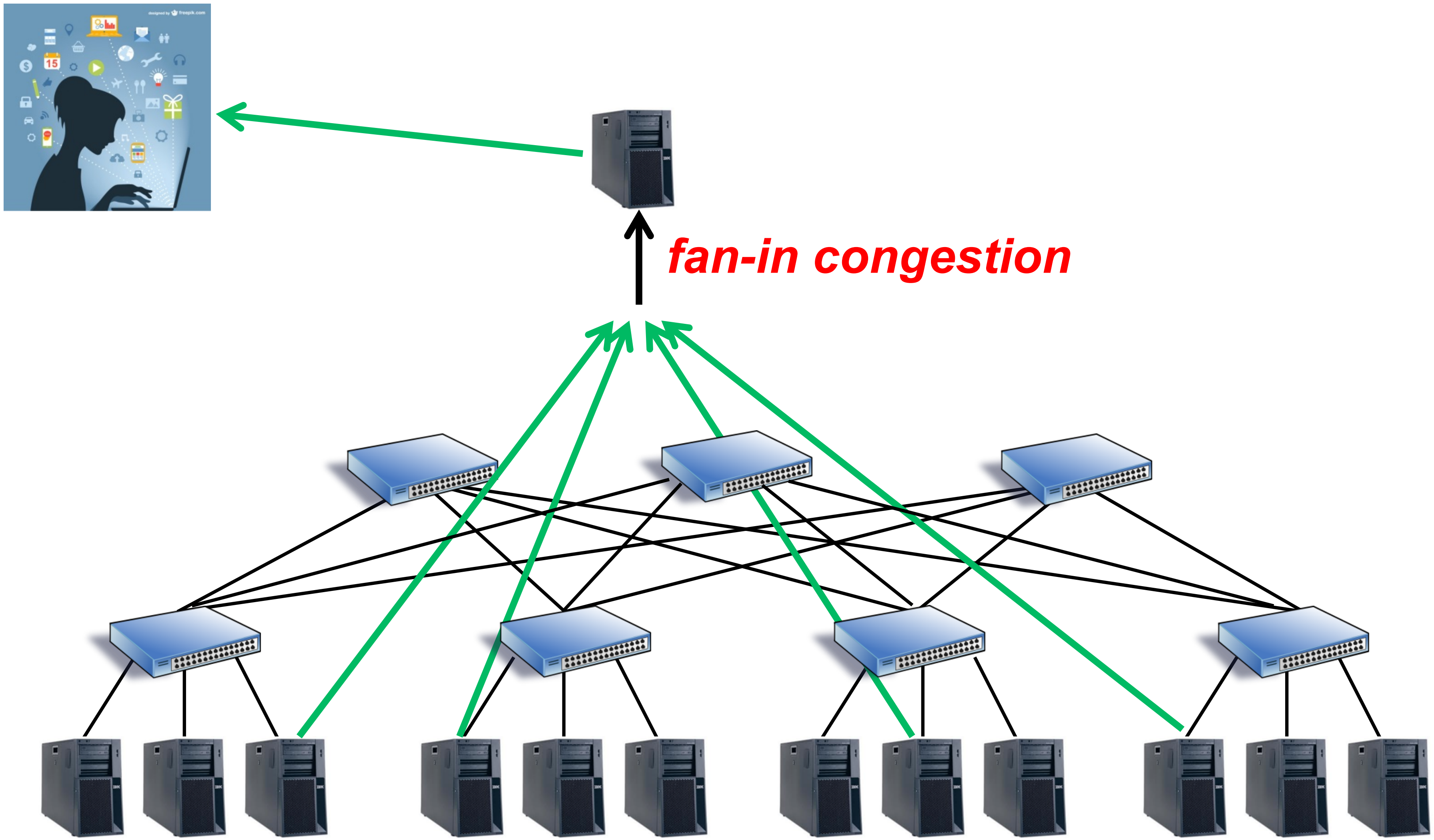
User is best satisfied if all partial results arrive on time



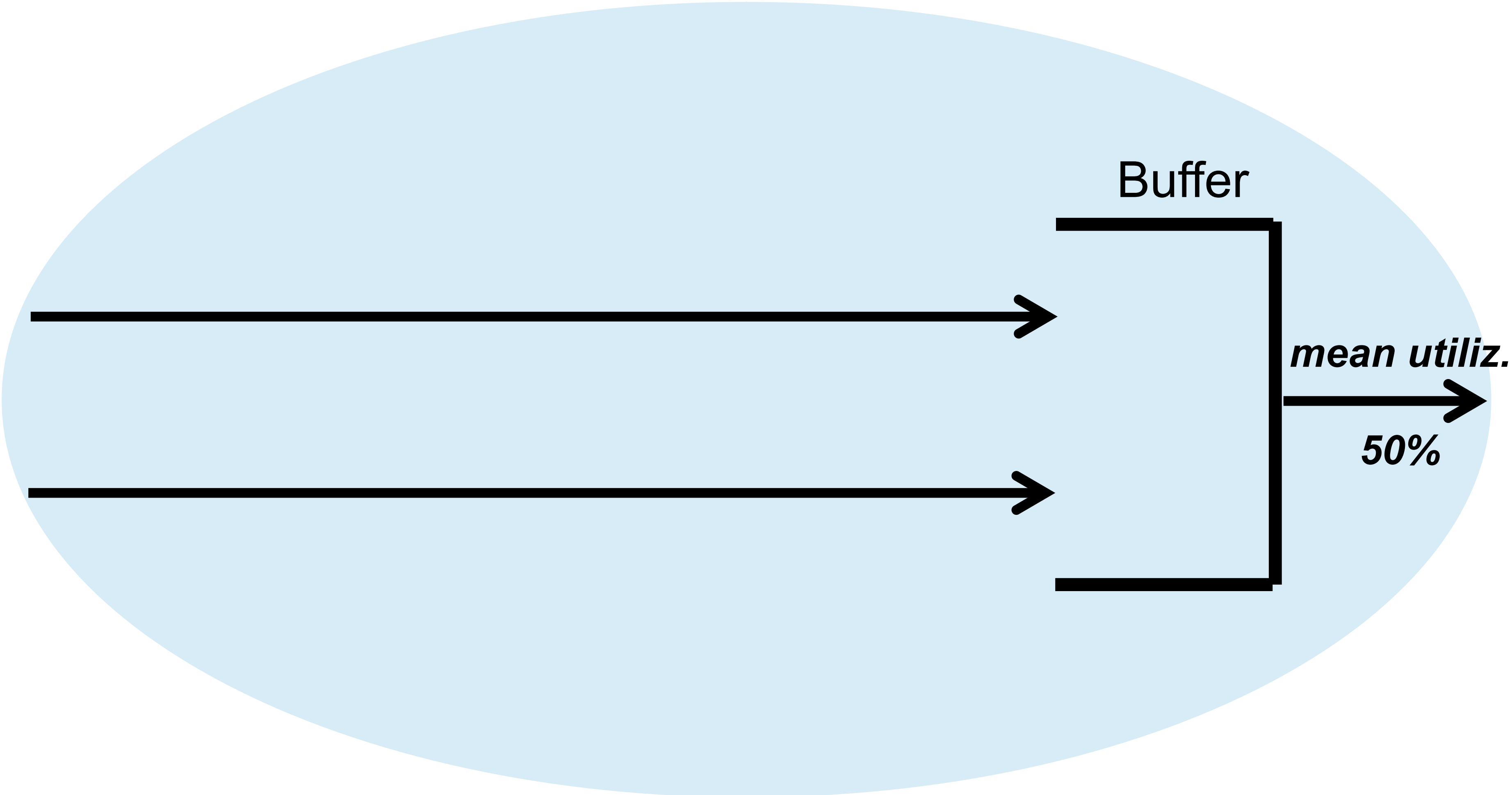
User is best satisfied if all partial results arrive on time



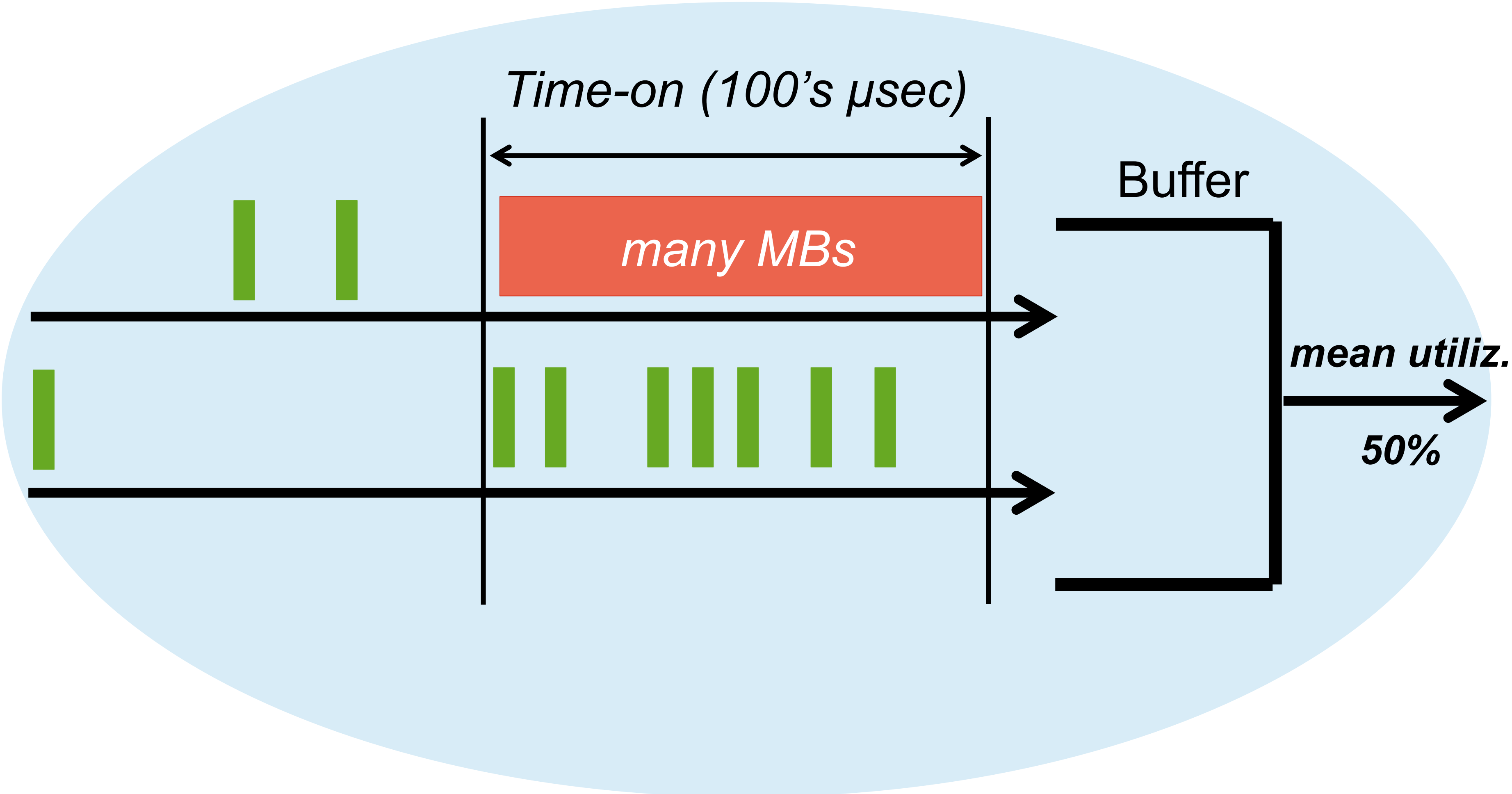
Compute close to data & flows are small ..but netw. suffers



“No problem: my network logs show 50% utilization”

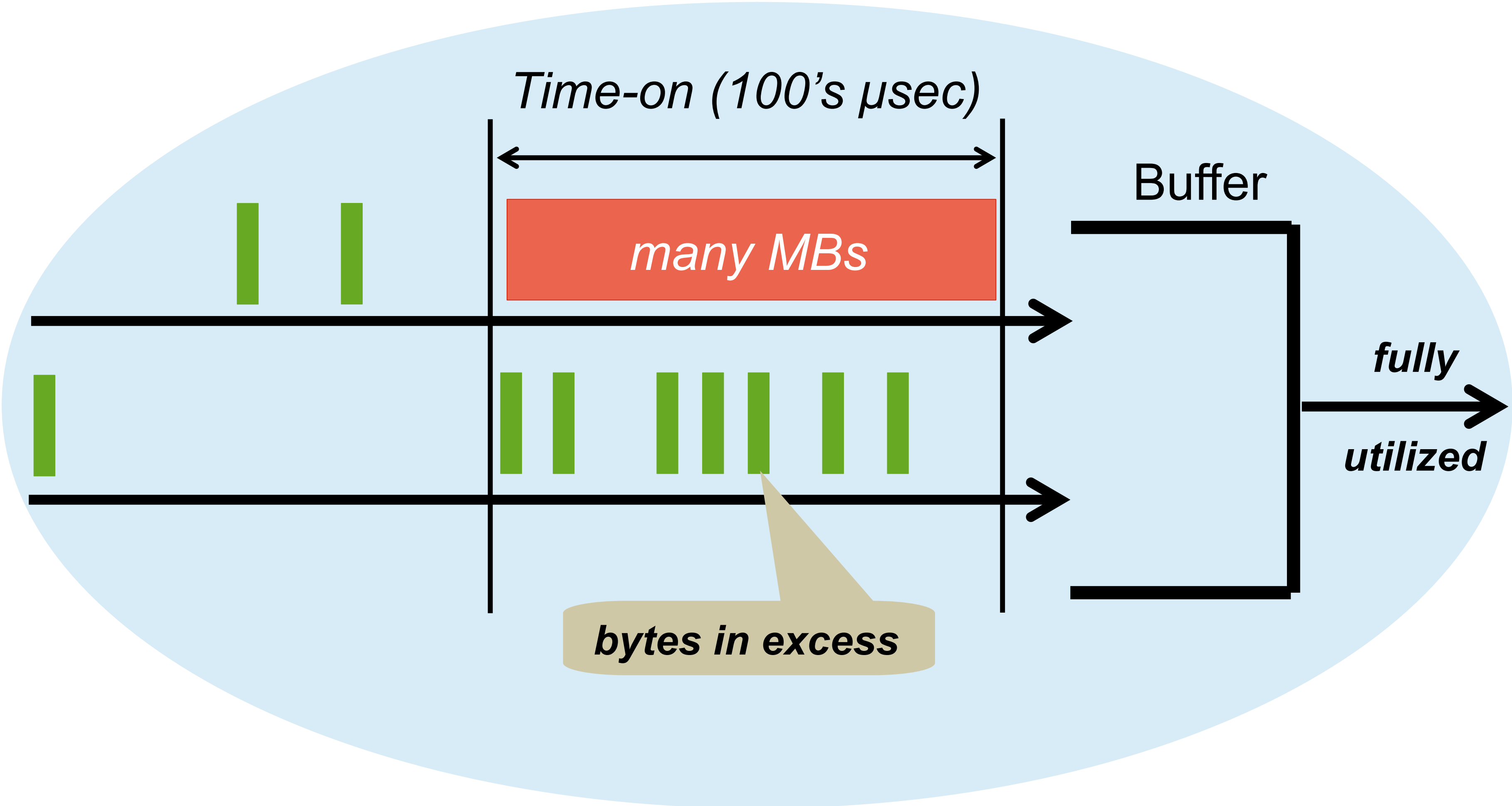


“No problem: my network logs show 50% utilization”



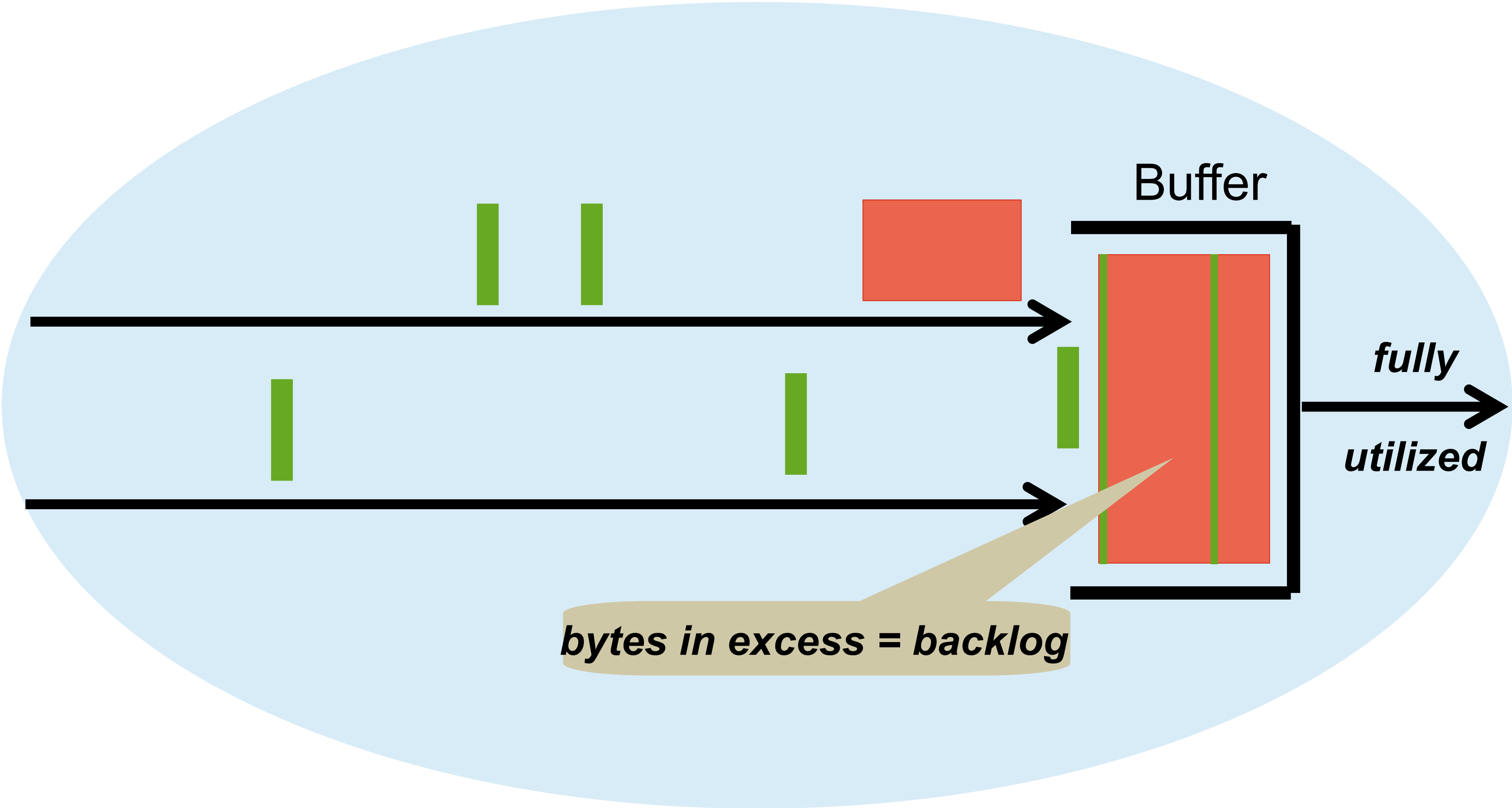
But bursty large flows & fan-in → strong fights @ short time scale

“No problem: my network logs show 50% utilization”



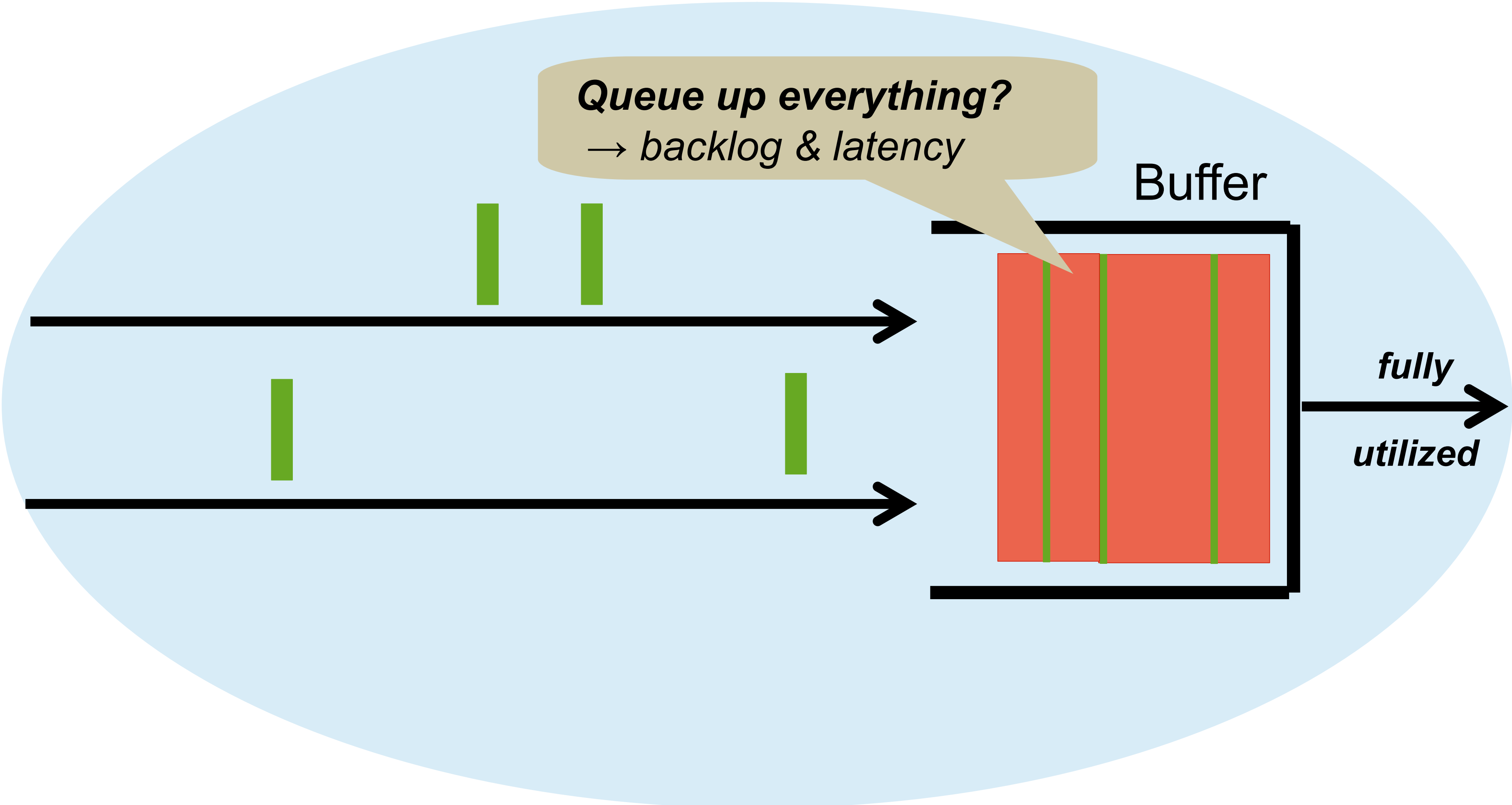
But bursty large flows & fan-in → strong fights @ short-time scale

Big problem: buffers will backlog

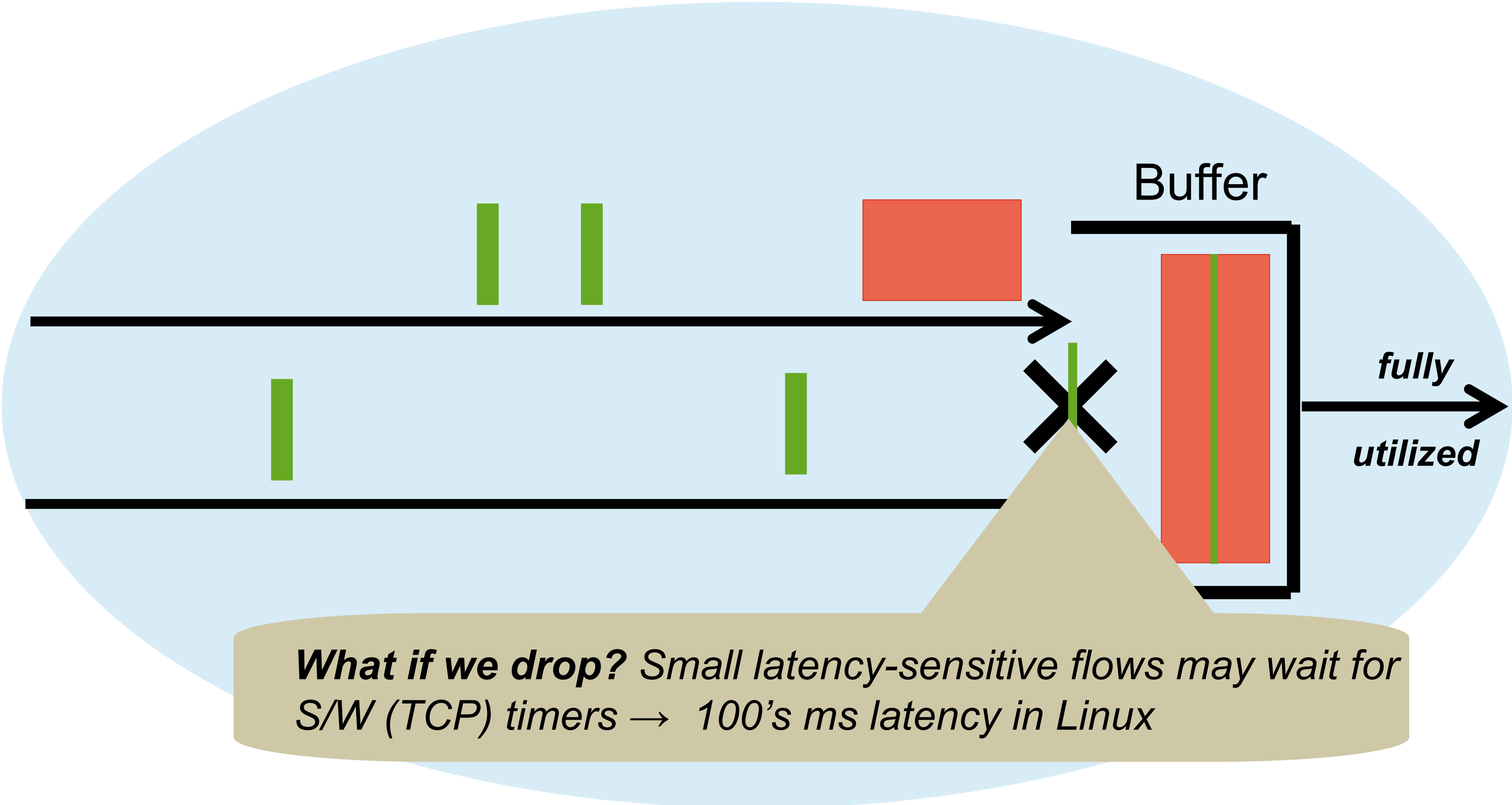


But bursty large flows & fan-in → strong fights @ short time scale

Big problem: buffers will backlog



Big problem: buffers will backlog



Outline

- Warehouse-scale datacenters and supercomputers
- Traffic characteristics in commercial datacenters
- **Efficient congestion control: an old, unresolved problem**
- Multipathing: benefits & issues
- RDMA: optimizing data copying
- Global virtual address space & routing

Network congestion: an old unresolved problem

- Effects
  - Unacceptable latencies
  - Productivity (throughput) reduction
- Need to live with it (hard) … or take measures

Network congestion: an old unresolved problem

- Effects
  - Unacceptable latencies
  - Productivity (throughput) reduction
- Need to live with it (hard) … or take measures

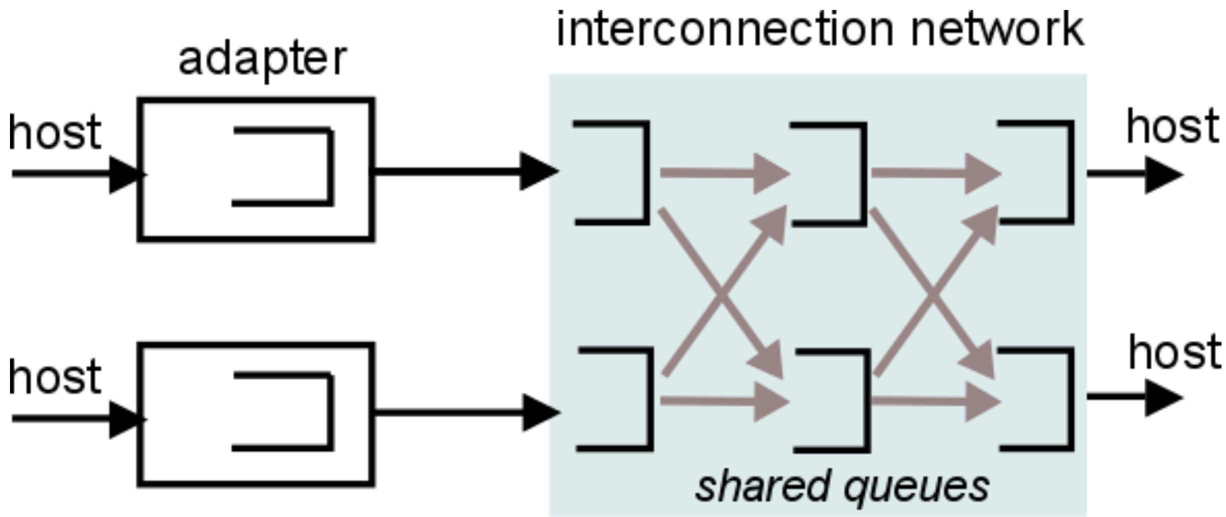
- TCP congestion control is universal … but not good enough
  - Conservative rate (window) control to ensure stability
    - converges to approximately fair rates after several RTTs
  - Cannot avoid backlogs & drops : recovery w. sluggish S/W retransmissions
    - bad for latency-critical flows
  - Long lat. (10's μsec) & many copies at hosts

Network congestion: an old unresolved problem

- Effects
  - Unacceptable latencies
  - Productivity (throughput) reduction
- Need to live with it (hard)
  … or take measures

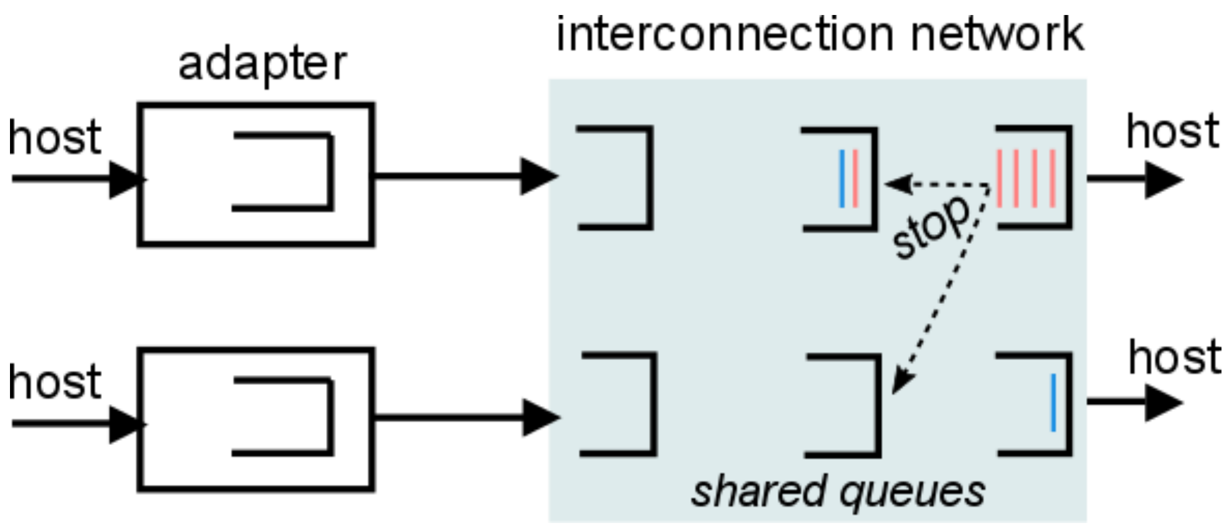
- TCP congestion control is universal … but not good enough
  - Conservative rate (window) control to ensure stability
    - converges to approximately fair rates after several RTTs
  - Cannot avoid backlogs & drops : recovery w. sluggish S/W retransmissions
    - bad for latency-critical flows
  - Long lat. (10's μsec) & many copies at hosts: ***can do much better w. RDMA***

Multistage interconnection network : abstract view



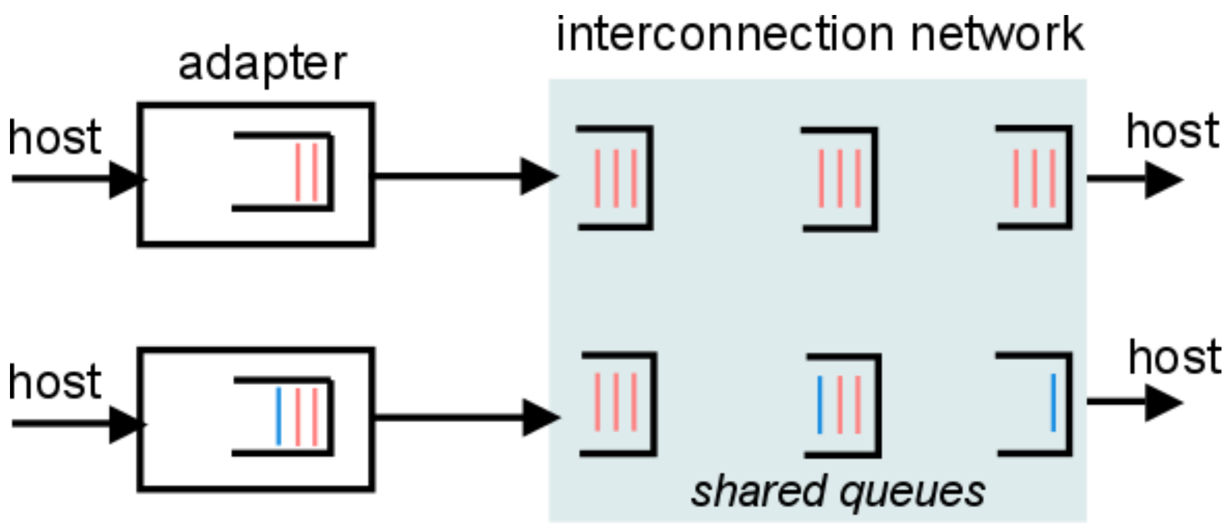
- One shared queue per switching element
- Multiple paths to destination

First measure: link-level flow control to avoid packet drops



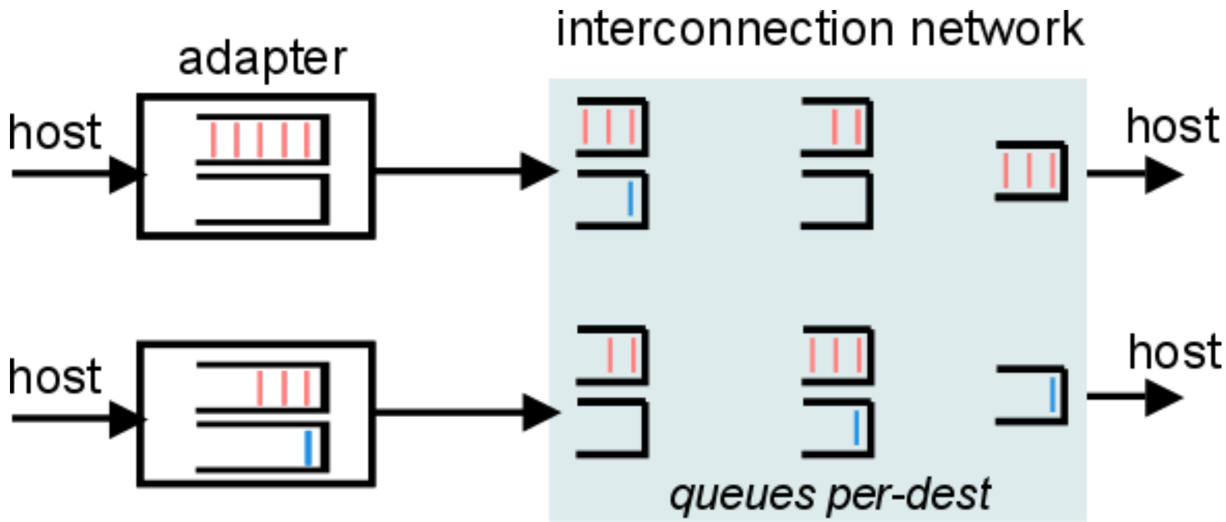
- Packets held in upstream buffer when downstream buffer full
  - Lossless networks: CEE (pause-based), Infiniband (credit-based)

First measure: link-level flow control to avoid packet drops



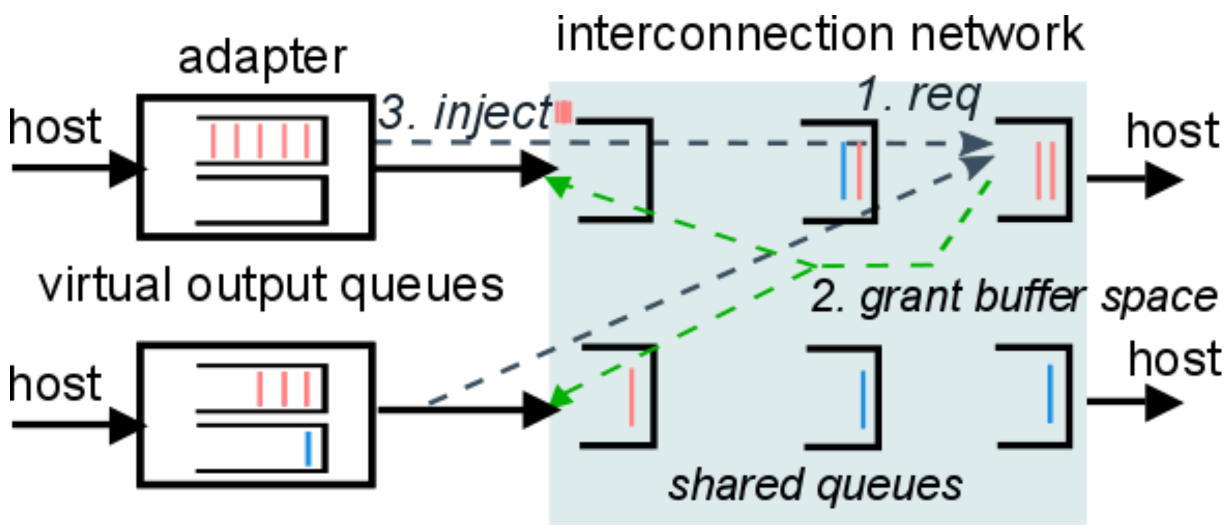
- Packets held in upstream buffer when downstream buffer full
  – Lossless networks: CEE (pause-based), Infiniband (credit-based)
- New problems
  – HOL blocking: nobody in Q can move because head packet is blocked
  – *Many* network buffers fill up not only at hotspot
    - Bad for latency-critical flows

A "high-end" accompanying measure: per-destination Qs



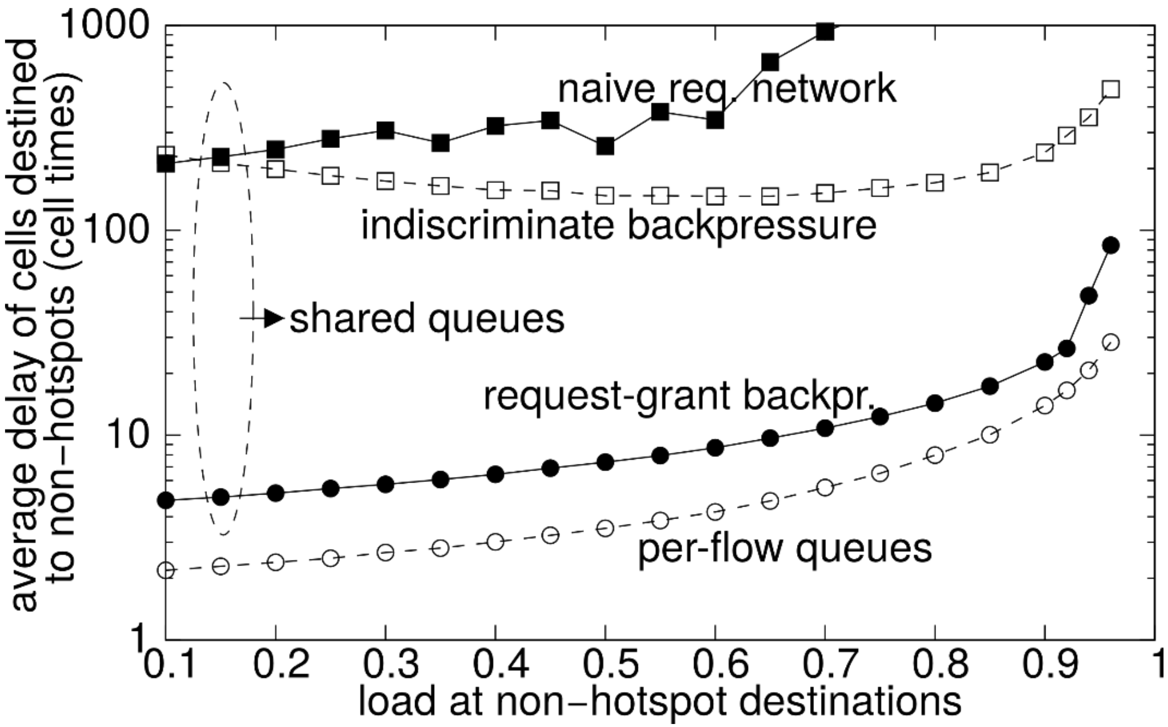
- Separate queues inside the network for every destination (Katevenis 1987, Sapunjis & Katevenis 2003)
  - Perfect isolation if link-level flow control stops only the queues of misbehaving flows (destinations)
  - Non-congested flows progress at full speed
- But cost of switching elements grows with network size

A sometimes affordable alternative: buffer reservations



- At high loads, reserve network buffer resources before injecting packets (Chrysos & Katevenis 2006, IBM 100G server-rack fabric 2014)
  - Packets wait at virtual output (per-dest) queues in front of network
- Shared in-fabric queues never exert backpressure
  - No HOL blocking and no backlogs
- But complicated schedulers & per-dest request queues (counters)

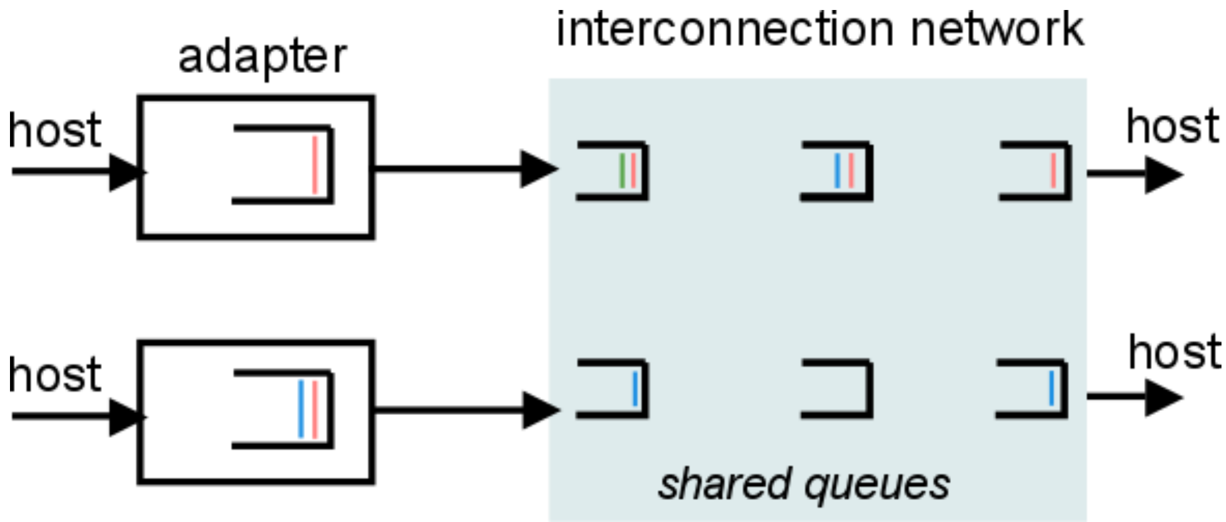
Simple flow control vs. per-flow queues vs. req-grant



Delay of innocent (non-congested) packets w. 1 congested dest

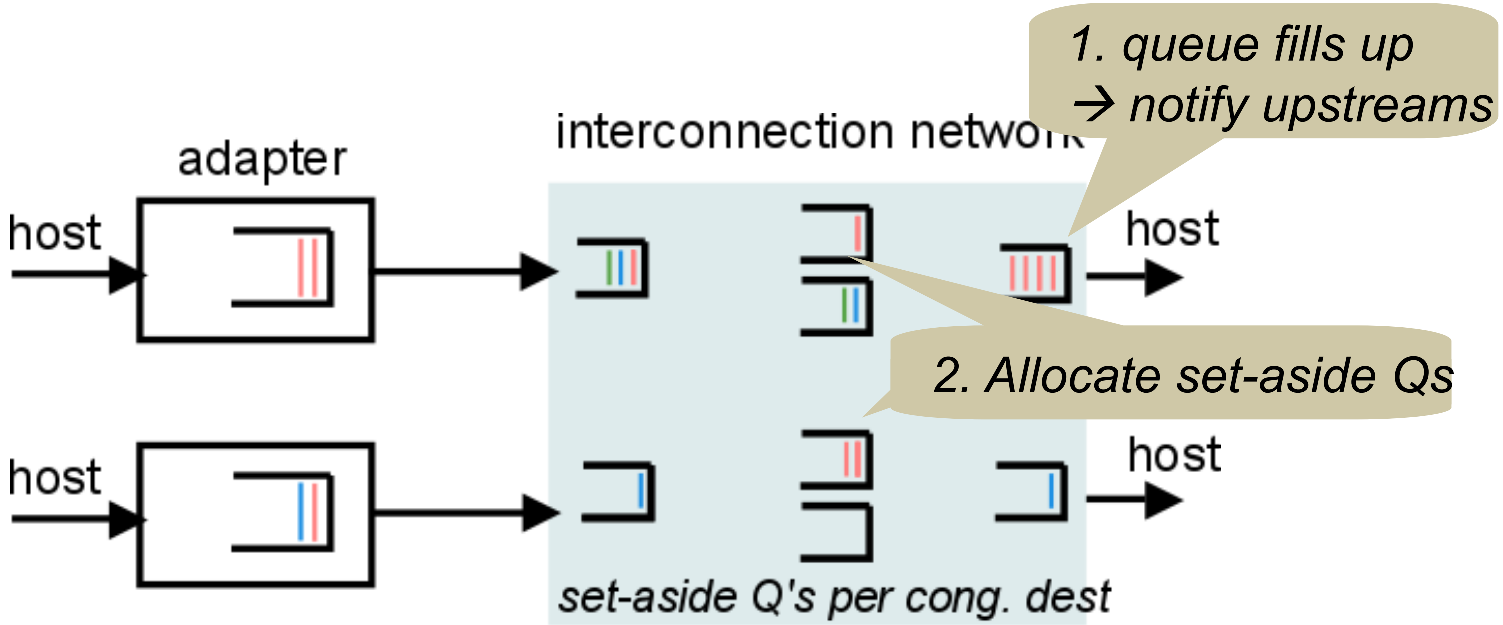
- Request-grant backpr. & per-flow queues best performance
  - small latencies for innocent cells – no backlogs & no HOL blocking in
- Simple flow control (indiscr. bkpr.) → innocent flows suffer
  - ~ 2 orders of magnitude higher latency

Dynamically allocated (per-flow) queues



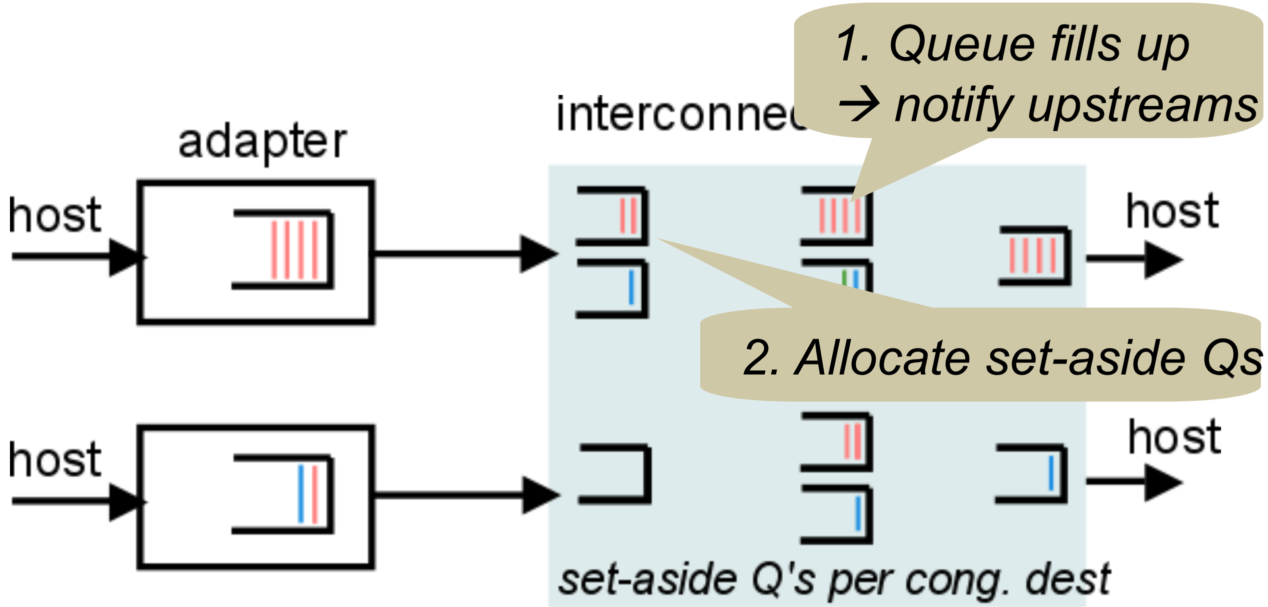
- Regional Explicit Congestion Notification (RECN – *Duato e.a.* HPCA 2005)
- Key insights:
  – Under normal conditions: one queue per link suffices

Dynamically allocated (per-flow) queues



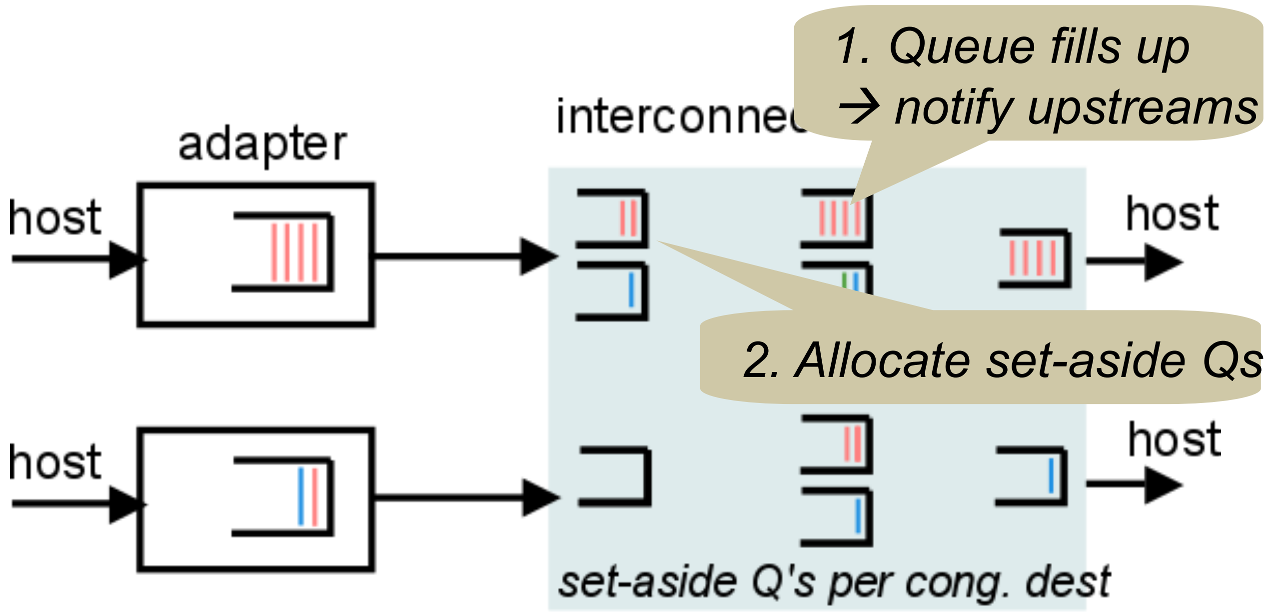
- Regional Explicit Congestion Notification (RECN – *Duato e.a.* HPCA 2005)
- Key insights:
  - Under normal conditions: one queue per link suffices
  - When a link becomes congested (queue fills up): notify upstreams to allocate private (set-aside) queues for traffic headed to congested link

Dynamically allocated (per-flow) queues



- Regional Explicit Congestion Notification (RECN – *Duato e.a.* HPCA 2005)
- Key insights:
  - Under normal conditions: one queue per link suffices
  - When a link becomes congested (queue fills up): notify upstreams to allocate private (set-aside) queues for traffic headed to congested link

Dynamically allocated (per-flow) queues



- Regional Explicit Congestion Notification (RECN – *Duato e.a.* HPCA 2005)
- Key insights:
  - Under normal conditions: one queue per link suffices
  - When a link becomes congested (queue fills up): notify upstreams to allocate private (set-aside) queues for traffic headed to congested link
- Complex link-level ctrl, costly for many concurrent hotspots

Industry-standard solutions

- Quantized Congestion Notification (QCN) for lossless (Converged Enhanced) Ethernet
  - – Congestion points @ netw. links send congestion notifications to sources: → allocate separate queues & rate control flows injections
    - Multiplicative decrease, additive increase ~ TCP
  - – Unfair and complex (*Chrysos e.a.* 2014)

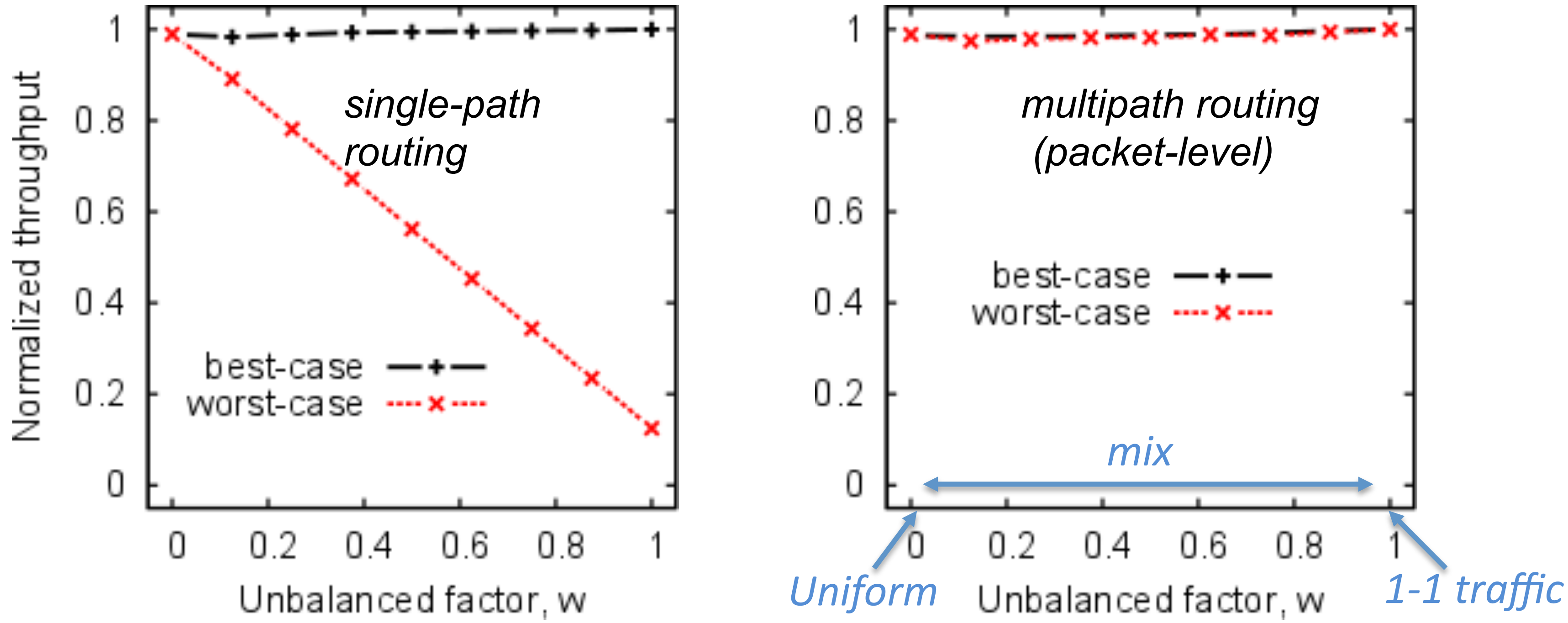
→ Most Ethernet networks are lossy and rely on TCP

- Infiniband congestion control
  - – very hard to tune and stabilize (Gusat *et al.* 2005)
  - – deployment levels: unknown

Outline

- Warehouse-scale datacenters and supercomputers
- Traffic characteristics in commercial datacenters
- Efficient congestion control: an old, unresolved problem
- **Multipathing: benefits & issues**
- RDMA: optimizing data copying
- Global virtual address space & routing

Multipath routing allows to exploit all available capacity

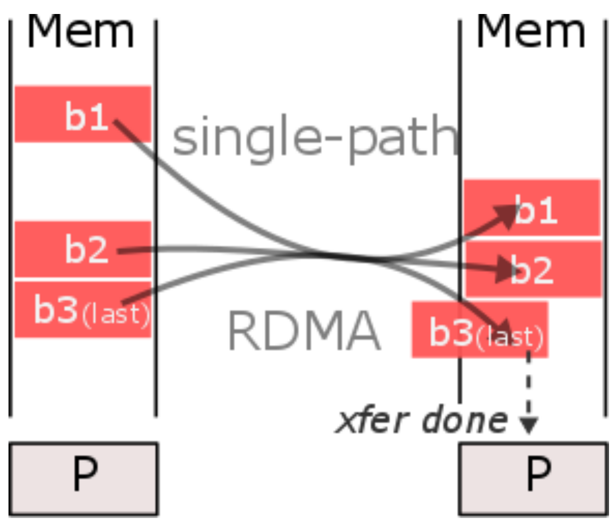


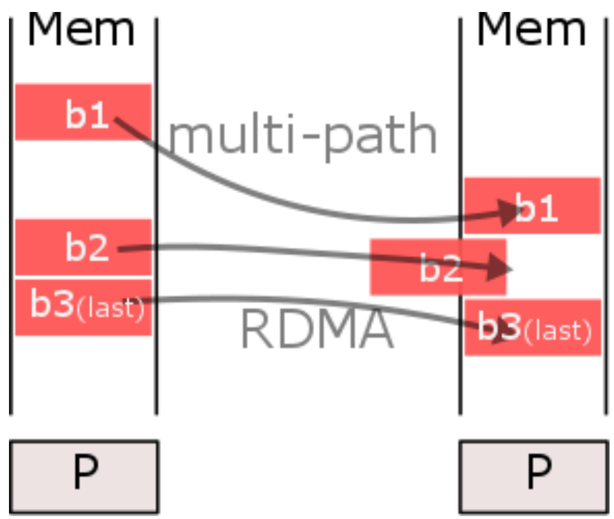
figures test different permutations on full-bisection-BW fat-tree

- Single-path routing: performance varies with spatial orientation of traffic
  - Same happens with industry-standard flow-level multipathing (ECMP routing)
- Packet-level multipathing consistently delivers full throughput

- Multipath Routing is also useful for Resilience

Multipath routing, however, complicates RDMA





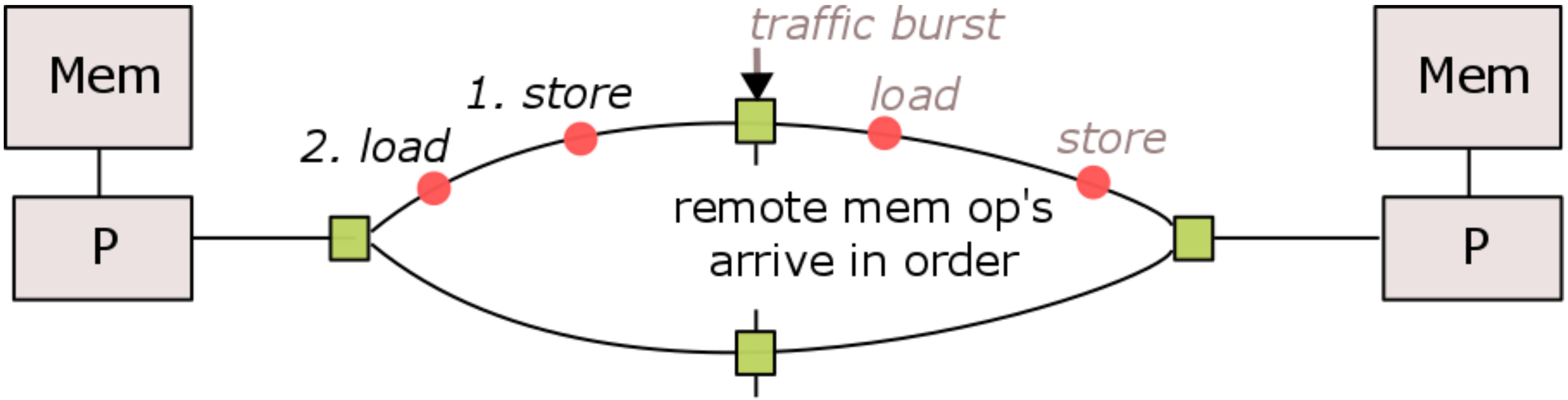
Single-path RDMA

- When dest gets last RDMA packet, it can wake up the processor to pick up the data

Multi-path RDMA

- RDMA packets may arrive out-of-order
  – Need other mechanisms to detect xfer completion

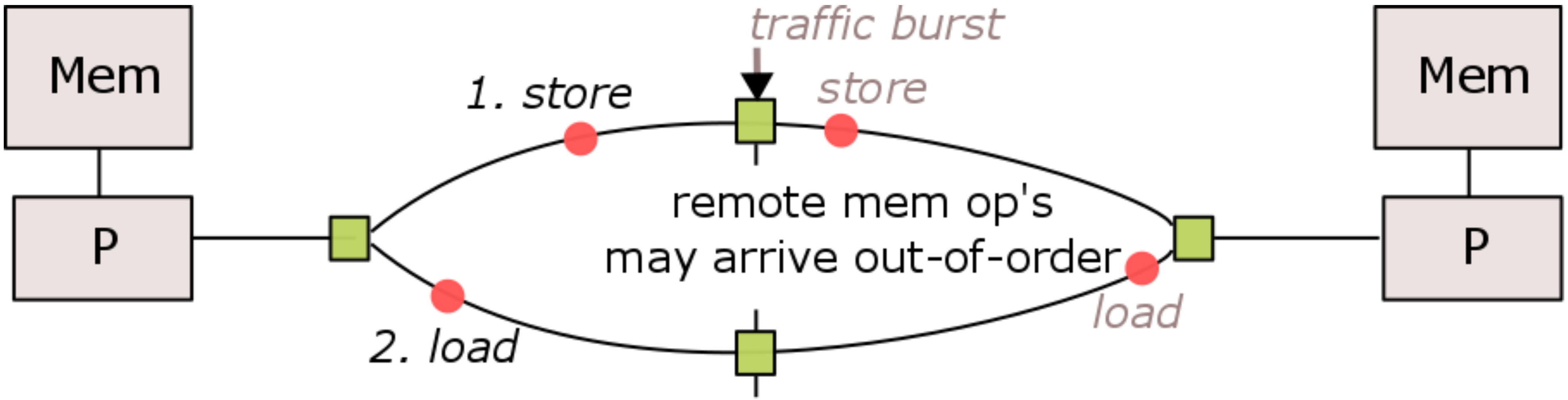
Multipathing also complicates remote memory operations



Remote store & load cmds for one (remote) memory location

- Semantically: “Load” should read what “Store” wrote
  - – OK with single-path routing

Multipathing also complicates remote memory operations



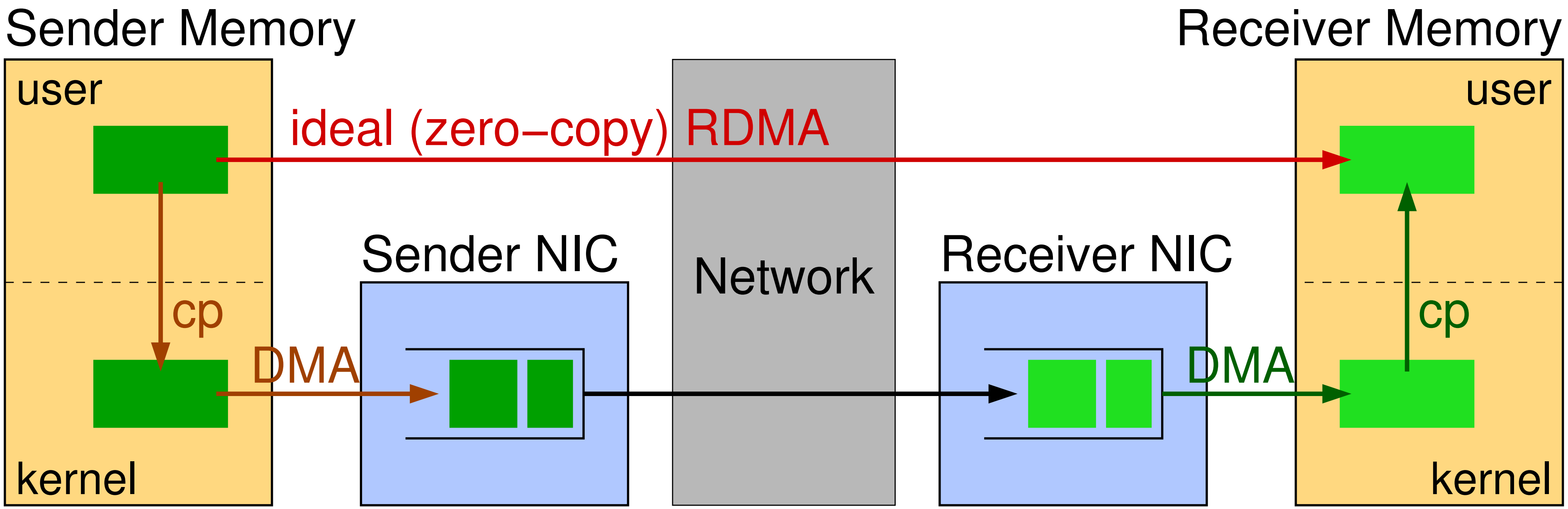
Remote store & load cmds for one (remote) memory location

- Semantically: "Load" should read what "Store" wrote
  - OK with single-path routing
  - Not necessarily true with multipath routing

Outline

- Warehouse-scale datacenters and supercomputers
- Traffic characteristics in commercial datacenters
- Efficient congestion control: an old, unresolved problem
- Multipathing: benefits & issues
- **RDMA: optimizing data copying**
- Global virtual address space & routing

Inefficiencies of traditional "Send" – *how to overcome them*



Up to 5x (!) inefficiencies

- Receiver copy NIC→User – *rcv addresses visible to sender: **R**DMA – PGAS*
- Protection – *virtualized, user-level DMA initiation, IOMMU*
- Buffer Pinning for DMA – *allow RDMA to fail, like page faults for ld/st*
- Send before receive buffer allocated – *fix the API / Application*
- Send buffer reuse immediately after send – *fix the API / Application*

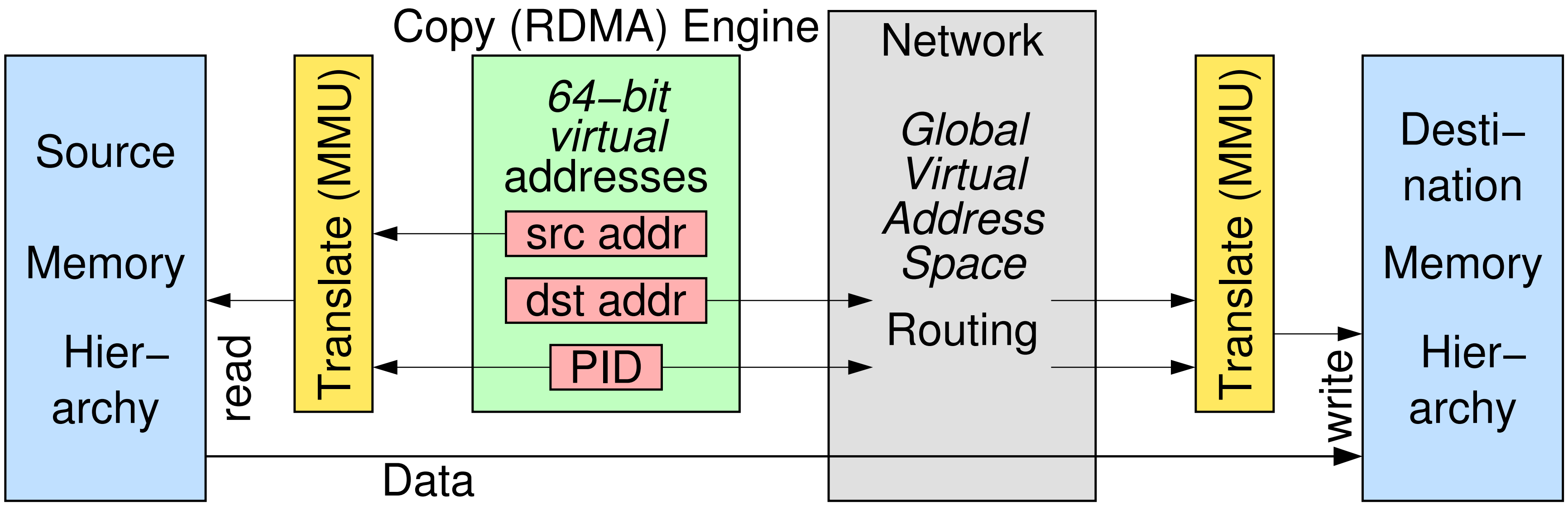
Wish List for an ideal Copy (RDMA) Engine

- User-Level RDMA Initiation:
  - Arguments to be full, arbitrary 64-bit *Virtual* Addresses
  - Control Registers to be virtualized and protected *per-process*
- No System Call necessary:
  - Virtual to Physical Address Translation via *HW MMU's* −not OS
  - Notification of Compl'n-Arrival: *per-process* Mailbox, not interrupt
- (true) Zero-Copy:
  - Any user page as source / destination
  - No need for pinning the src-dst pages in-memory: allow for translation failures during RDMA operation, resuting in notification of incomplete operation −like normal page-faults
  - Also useful for **Resilience**
- Exascale Global Addr. Space: full 64-b virtual addr. (+PID) throughout
- Performance: multi-channel engine; per-channel flow/rate control

Outline

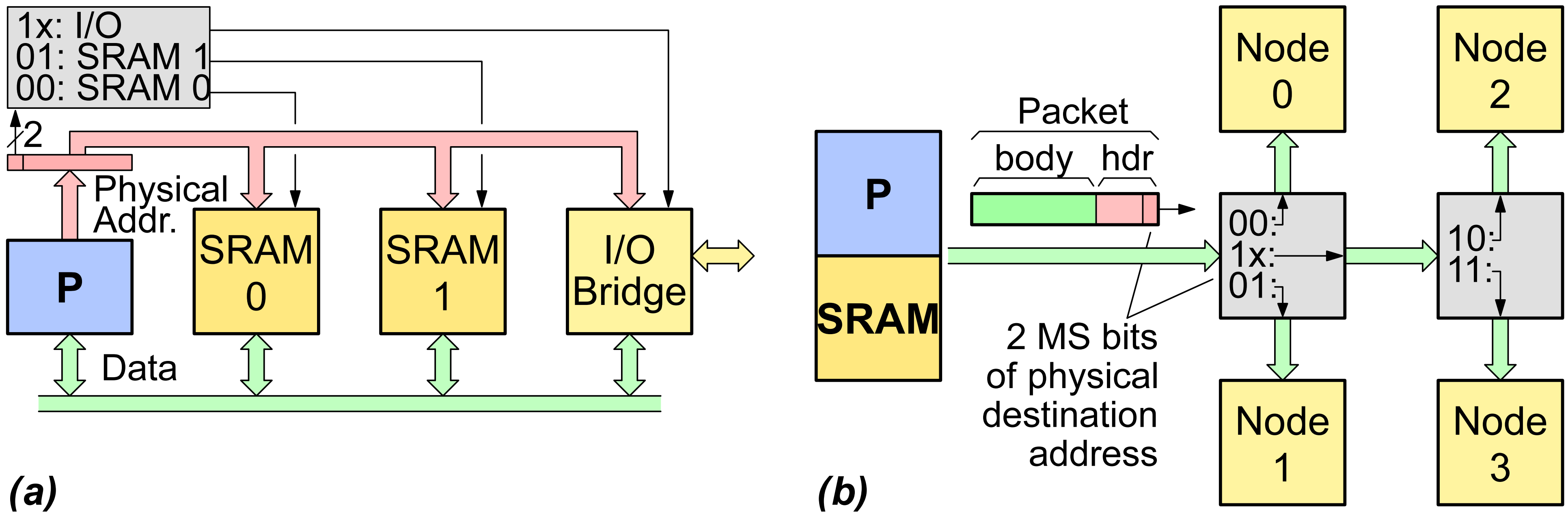
- Warehouse-scale datacenters and supercomputers
- Traffic characteristics in commercial datacenters
- Efficient congestion control: an old, unresolved problem
- Multipathing: benefits & issues
- RDMA: optimizing data copying
- **Global virtual address space & routing**

User-Level Commun. in a true Global Virtual Address Space



- GVAS for exascale needs 64-bit addresses, with (global) protection domain identifier either incorporated in them or as extra bits
- Sophisticated network routing based on GVA will allow (large) page (segment) live migration –see "Progressive Address Translation " in Katevenis 2007 paper http://www.ics.forth.gr/carv/ipc/ldstgen_katevenis07.pdf

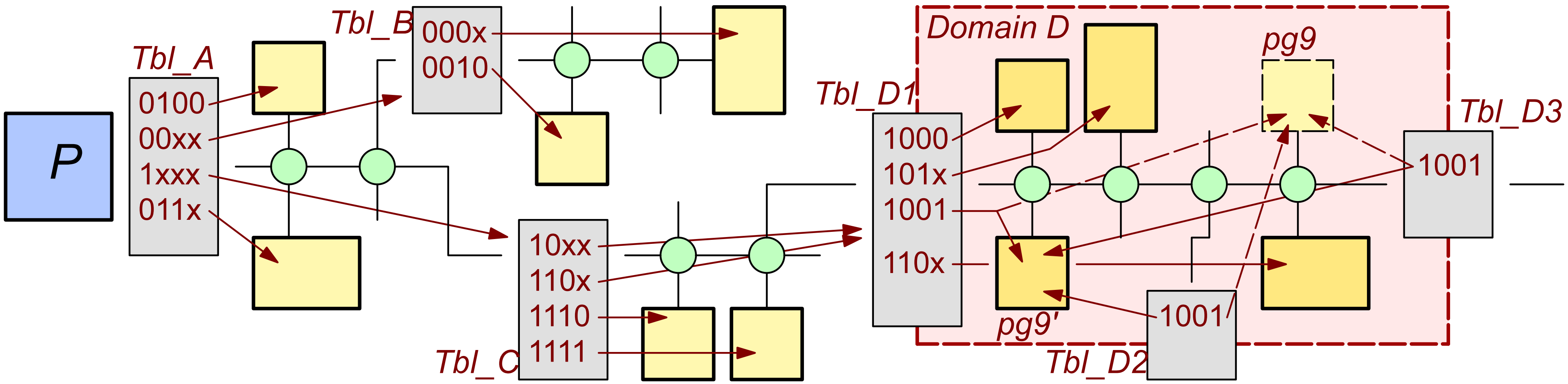
SARC Project (2005-09): Network Routing as Generalization of Address Decoding



- Physical Address Decoding in a uniprocessor
- Geographical Address Routing in a multiprocessor

http://www.ics.forth.gr/carv/ipc/ldstgen_katevenis07.pdf

Progressive Address Translation: Localize Migration Updates



- Packets carry global virtual addresses
- Tables provide physical route (address) for the next few steps
- When page 9 migrates within D, only tables in that domain need updating
- Variable-size-page translation tables look like internet routing tables (longest-prefix matches if we want small-page-within-big-region migration)
- Tables that partition the system, for protection against untrusted operating systems, look like internet firewalls

Conclusions

- Datacenter (and Supercomputer) Interconnects: increasingly important – challenges & opportunities
- Congestion Management: important, hard, unresolved
  - quick feedback, throttle sources, avoid drops, avoid deadlocks
- Multipathing:
  - good performance, useful for Resilience, but out-of-order delivery
- RDMA & Global Virt. Addr. Sp. for optimizing data copying:
  - known techniques, now need to convince industry to adopt them
- Routing: related to address translation and multipathing