Network Requirements for Resource Disaggregation

Peter X. Gao*, Akshay Narayan (MIT)*, Sagar Karandikar*, Joao Carreira*, Sangjin Han*
Rachit Agarwal (Cornell), Sylvia Ratnasamy, Scott Shenker
petergao,akshay,sagark,joao,sangjin,ragarwal,sylvia,shenker@eecs.berkeley.edu, UC Berkeley

Traditional datacenters are designed as a collection of servers, each of which tightly couples the resources required for computing tasks. Recent industry trends suggest a paradigm shift to a disaggregated datacenter (DDC) architecture containing a pool of resources, each built as a standalone resource blade, and interconnected using a network fabric. Examples include Facebook Disaggregated Rack [1], HP "The Machine" [2], Intel Rack Scale Architecture [3], SeaMicro [4] and Firebox [6].

A key enabling (or blocking) factor for disaggregation will be the network – to support good application-level performance it is critical that the network fabric provide low latency communication even under the increased traffic load that disaggregation introduces. Here, we use a workload-driven approach to derive the minimum latency and bandwidth requirements that the network in disaggregated datacenters must provide to avoid degrading application-level performance and explore the feasibility of meeting these requirements with existing system designs and commodity networking technology.

Using a combination of emulation, simulation, and implementation, we evaluate these minimum network requirements in the context of ten workloads spanning seven popular open-source systems — Hadoop, Spark, GraphLab, Timely dataflow, Spark Streaming, memcached, HERD, and SparkSQL. Our key findings are:

- Network bandwidth in the range of $40-100$Gbps is sufficient to maintain application-level performance within 5% of that in existing datacenters; this is easily in reach of existing switch and NIC hardware.
- Network latency in the range of $3-5\mu$s is needed to maintain application-level performance. This is a challenging task. Our analysis suggests that the primary latency bottleneck stems from network software rather than hardware: we find the latency introduced by the endpoint is roughly 66% of the inter-rack latency and roughly 81% of the intra-rack latency. Thus many of the switch hardware optimizations (such as terabit links) pursued today can optimize only a small fraction of the overall latency budget. Instead, work on bypassing the kernel for packet processing and on NIC integration could significantly enhance the feasibility of resource disaggregation.
- The root cause of the above bandwidth and latency requirements is the application's memory bandwidth demand.
- While most efforts focus on disaggregating at the rack scale, our results show that for many applications disaggregation at the datacenter scale is entirely feasible.
- Finally, our study shows that transport protocols frequently deployed in today's datacenters (TCP or DCTCP) fail to meet our target requirements for low latency communication with the DDC workloads. However, some recent research proposals [5, 8] do provide the necessary end-to-end latencies.

Taken together, our study suggests that resource disaggregation need not be gated on the availability of new networking hardware: instead, minimal performance degradation can be achieved with existing network hardware (either commodity, or available shortly). Please refer to [7] for details of our study.

References

[1] Facebook Disaggregated Rack. `http://goo.gl/6h2Ut`.

[2] HP The Machine. `http://goo.gl/wkDDEi`.

[3] Intel RSA. `http://goo.gl/f0U6Tp`.

[4] SeaMicro Technology. `http://goo.gl/vXpkMK`.

[5] M. Alizadeh, S. Yang, M. Sharif, S. Katti, N. McKeown, B. Prabhakar, and S. Shenker. pFabric: Minimal Near-optimal Datacenter Transport. SIGCOMM 2013.

[6] K. Asanović. FireBox: A Hardware Building Block for 2020 Warehouse-Scale Computers. FAST 2014.

[7] P. X. Gao, A. Narayan, S. Karandikar, J. Carreira, S. Han, R. Agarwal, S. Ratnasamy, and S. Shenker. Network Requirements for Resource Disaggregation. OSDI 2016.

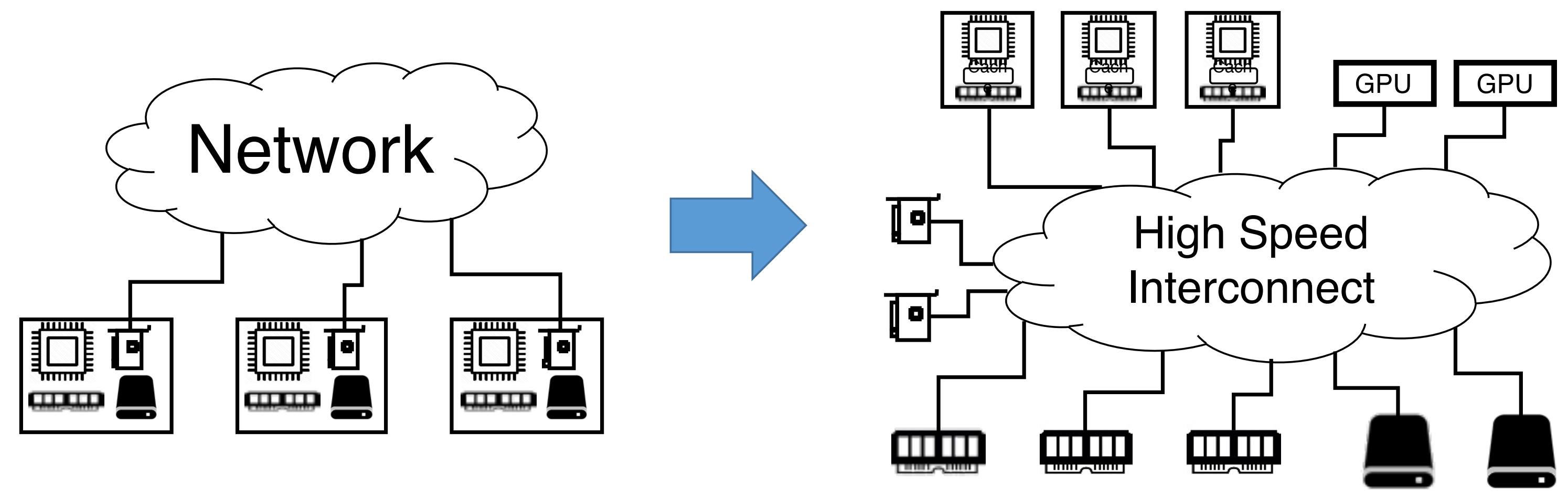
[8] P. X. Gao, A. Narayan, G. Kumar, R. Agarwal, S. Ratnasamy, and S. Shenker. pHost: Distributed Near-optimal Datacenter Transport Over Commodity Network Fabric. CoNEXT 2015.

*Student authors. No demo for this poster.

Network Requirements for Resource Disaggregation

Peter Gao, Akshay Narayan(MIT), Sagar Karandikar, Joao Carreira, Sangjin Han, Rachit Agarwal(Cornell), Sylvia Ratnasamy, Scott Shenker (UC Berkeley)

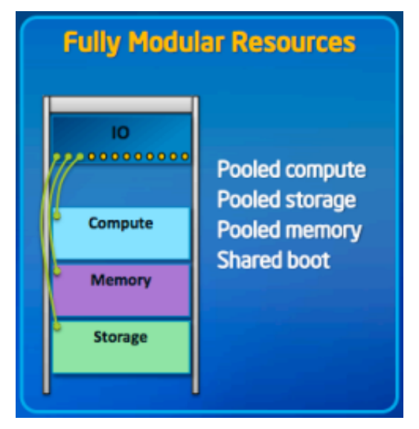
What is resource disaggregation?



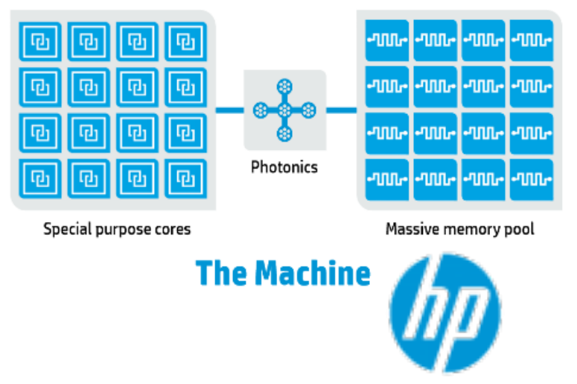
Server centric architecture: each server is a self-contained system comprise of CPU, memory, disk, and other peripherals

Disaggregated datacenter: each endpoint is a resource blade of a single type of resource, connected by high speed interconnect

Where is it happening?



Intel - RSA

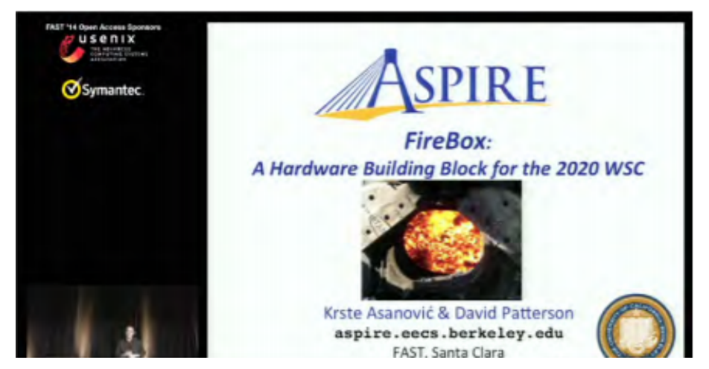


HP – The Machine

Facebook Disaggregated Rack

Huawei - NUWA

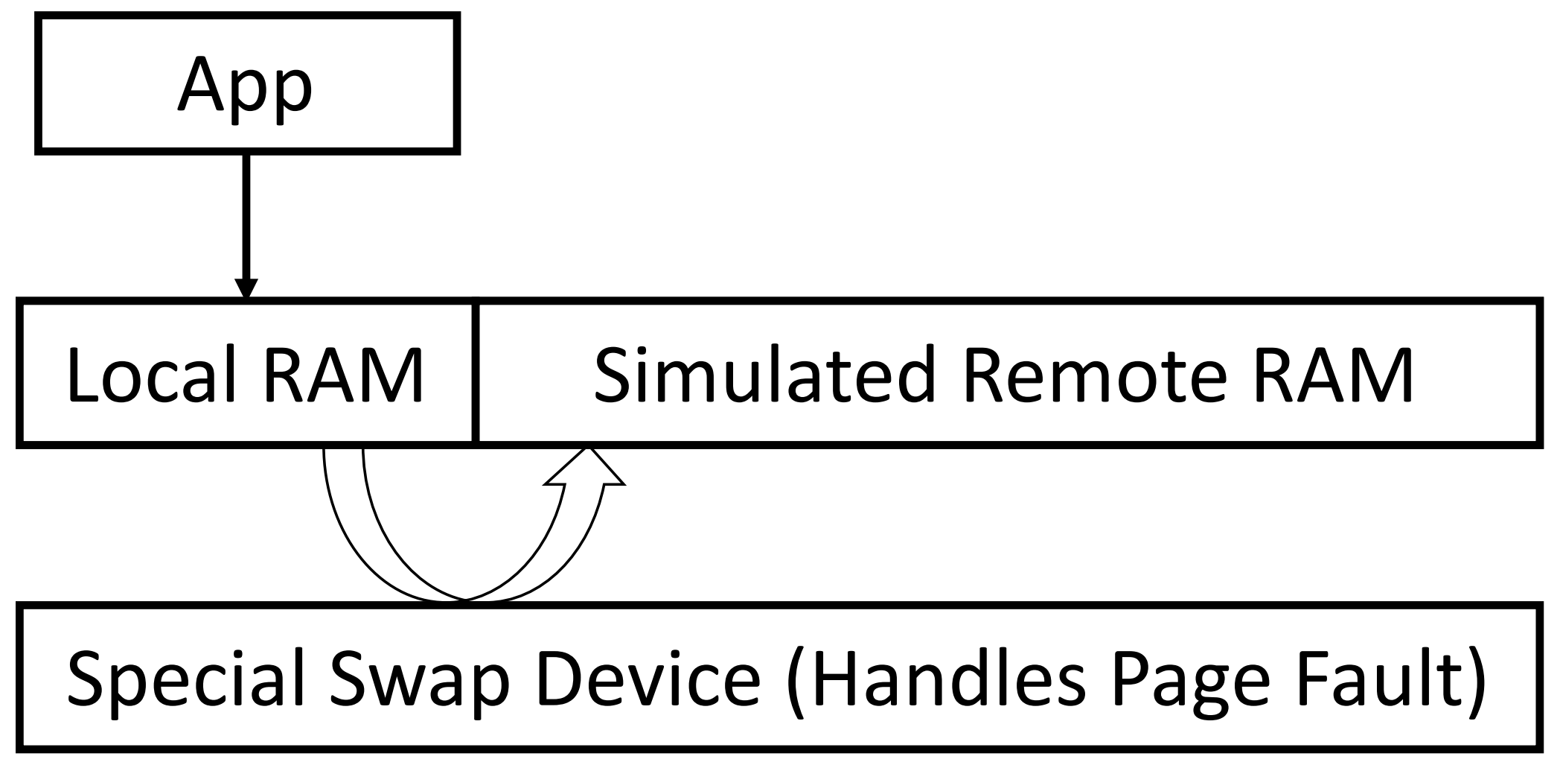
SeaMicro SM15000



Berkeley - Firebox

Goals and **Methodology**

- Latency and bandwidth requirement of the network
- Are current network designs sufficient?



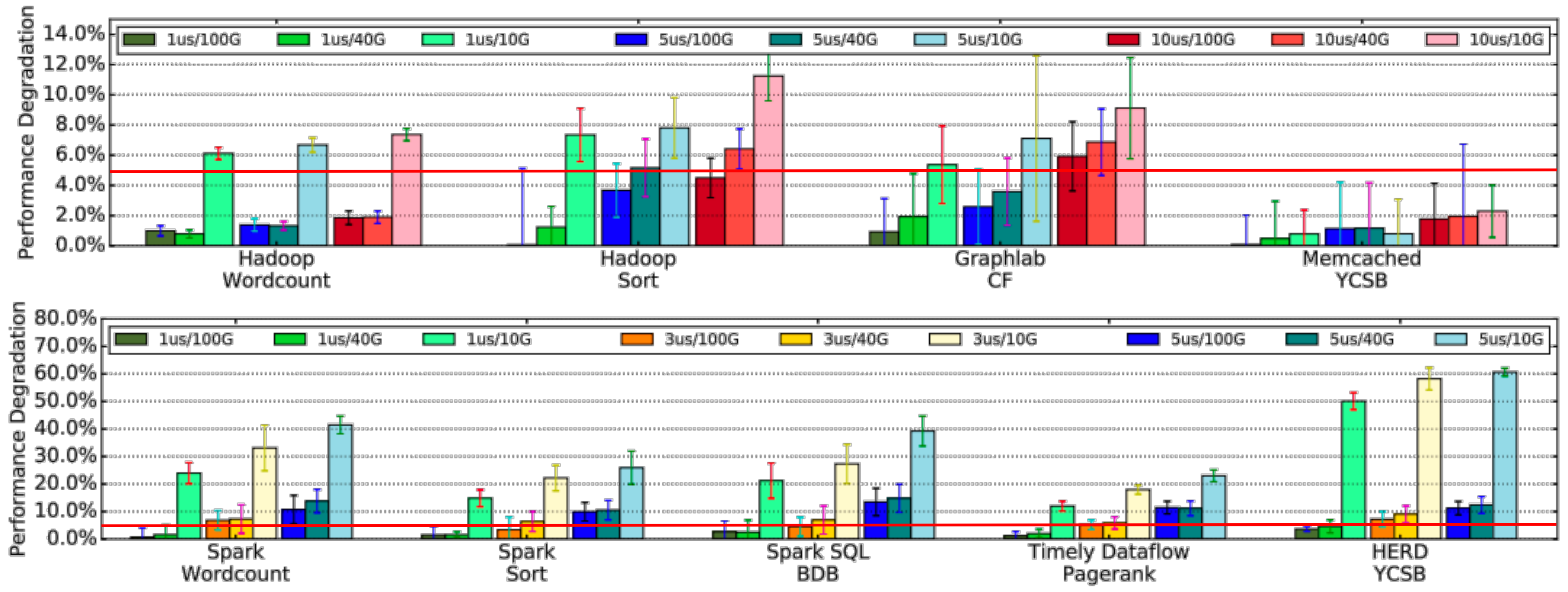
App

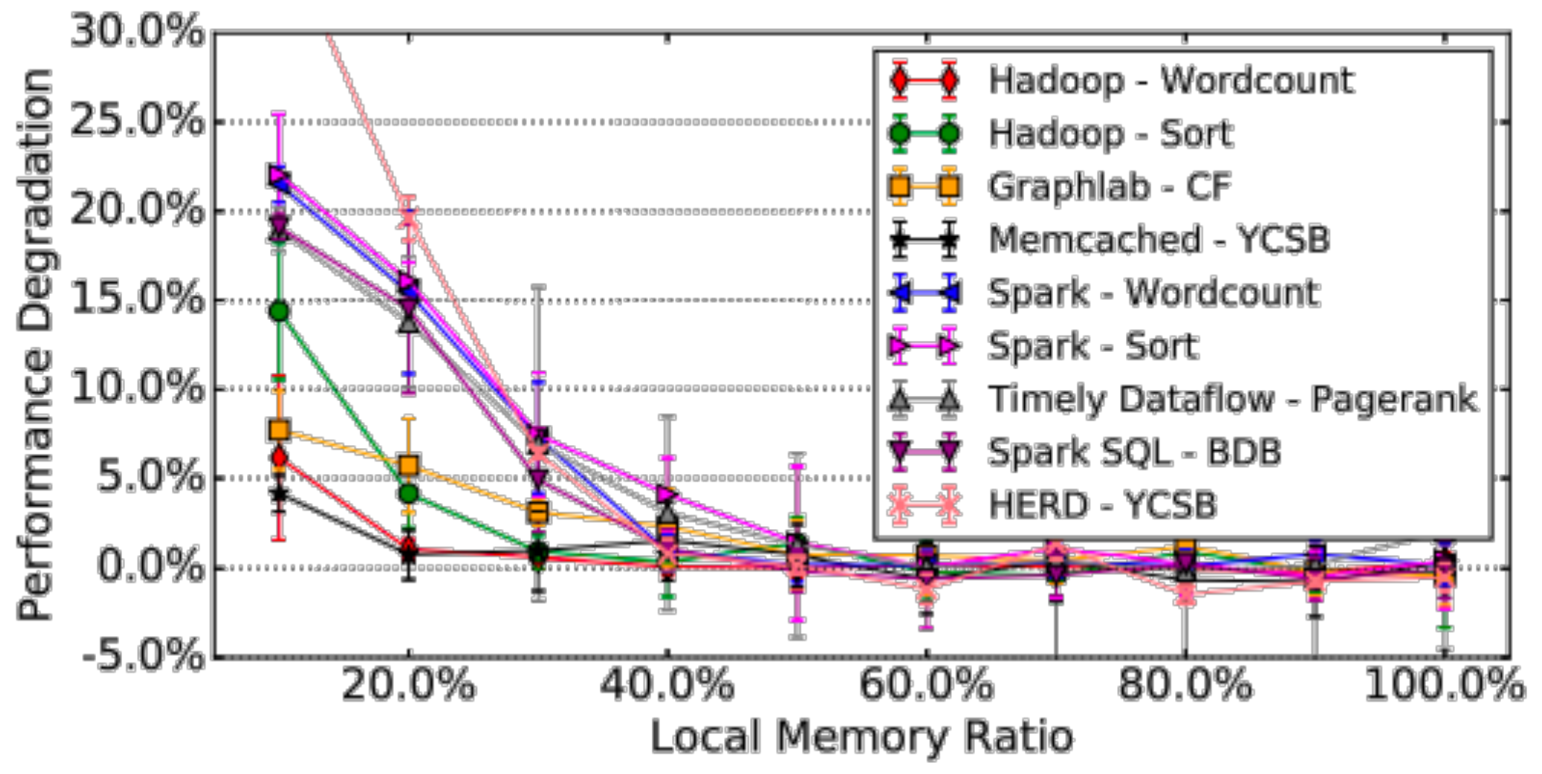
Local RAM | Simulated Remote RAM

Special Swap Device (Handles Page Fault)

- Special swap device capable of latency injection
- Run 10 workloads on 8 frameworks

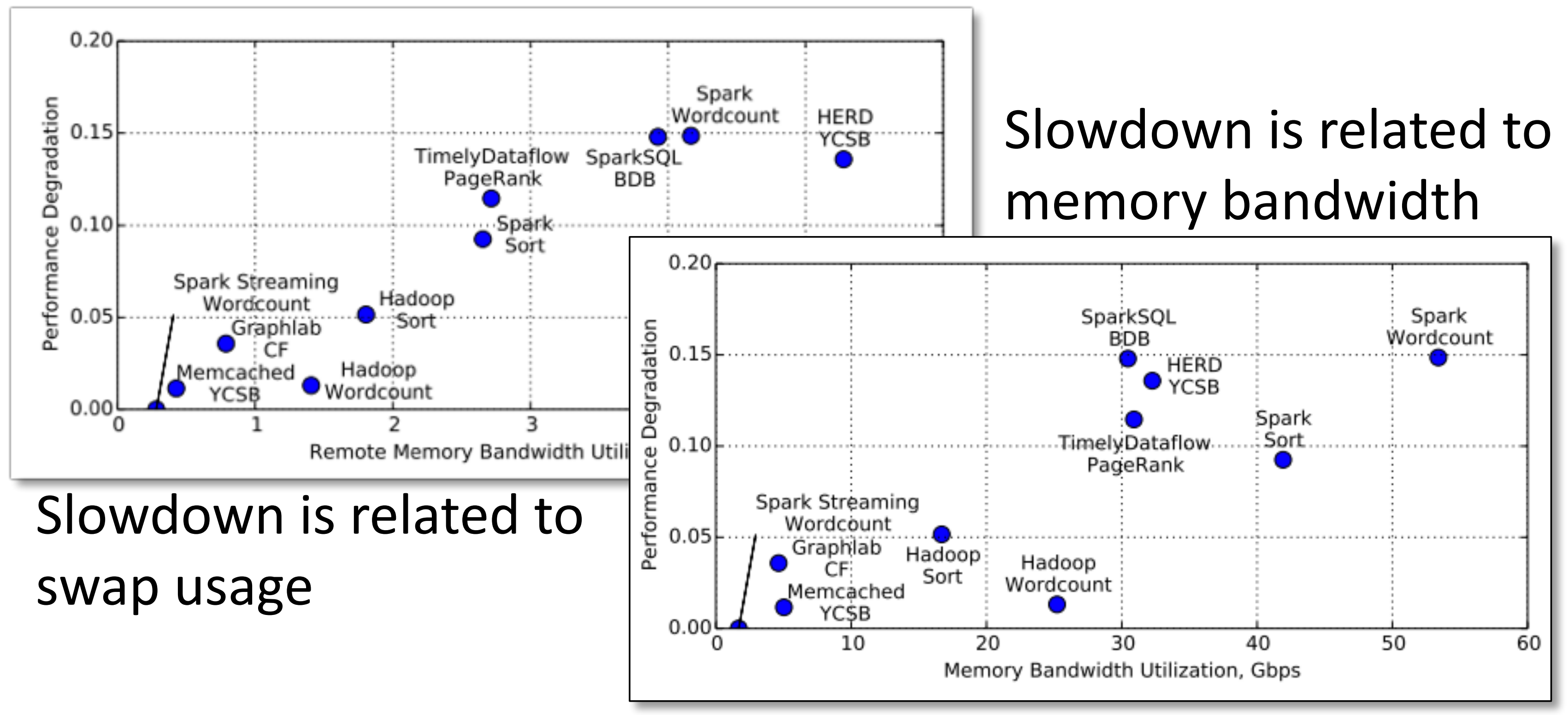
5% degradation for legacy apps





- 20% -30% local memory
- 40Gbps Bandwidth
- 1-5us latency

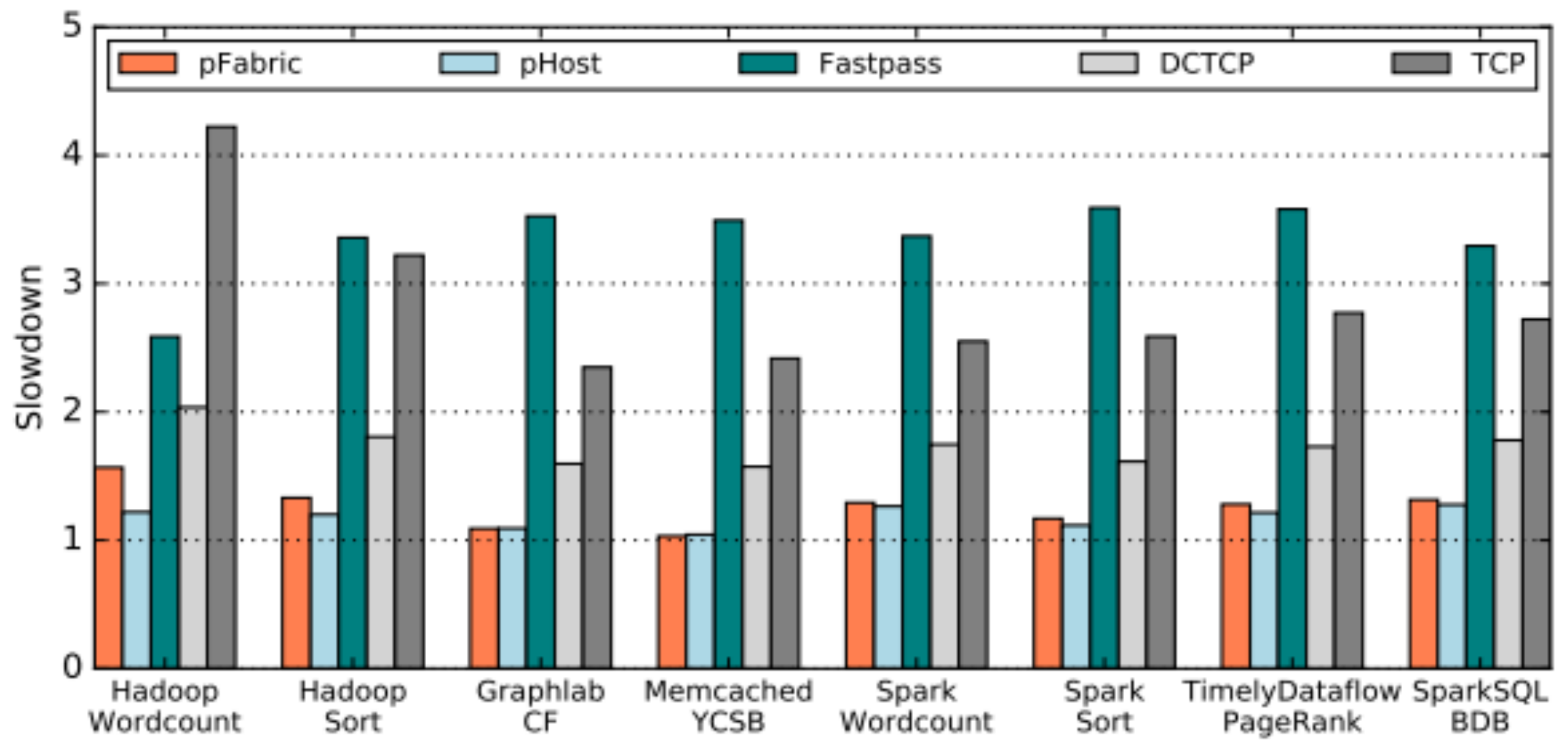
Understand Degradation



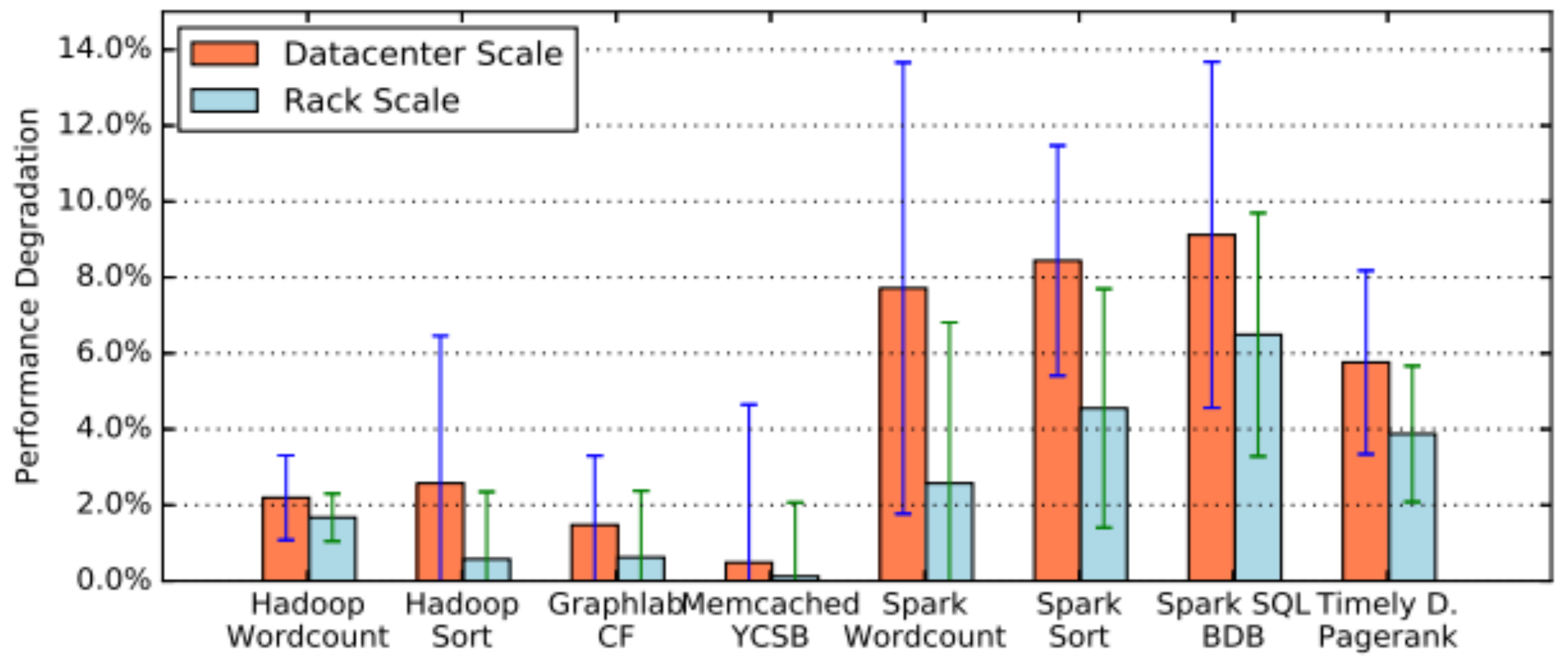
Slowdown is related to memory bandwidth

Slowdown is related to swap usage

Can existing **transport** handle this?

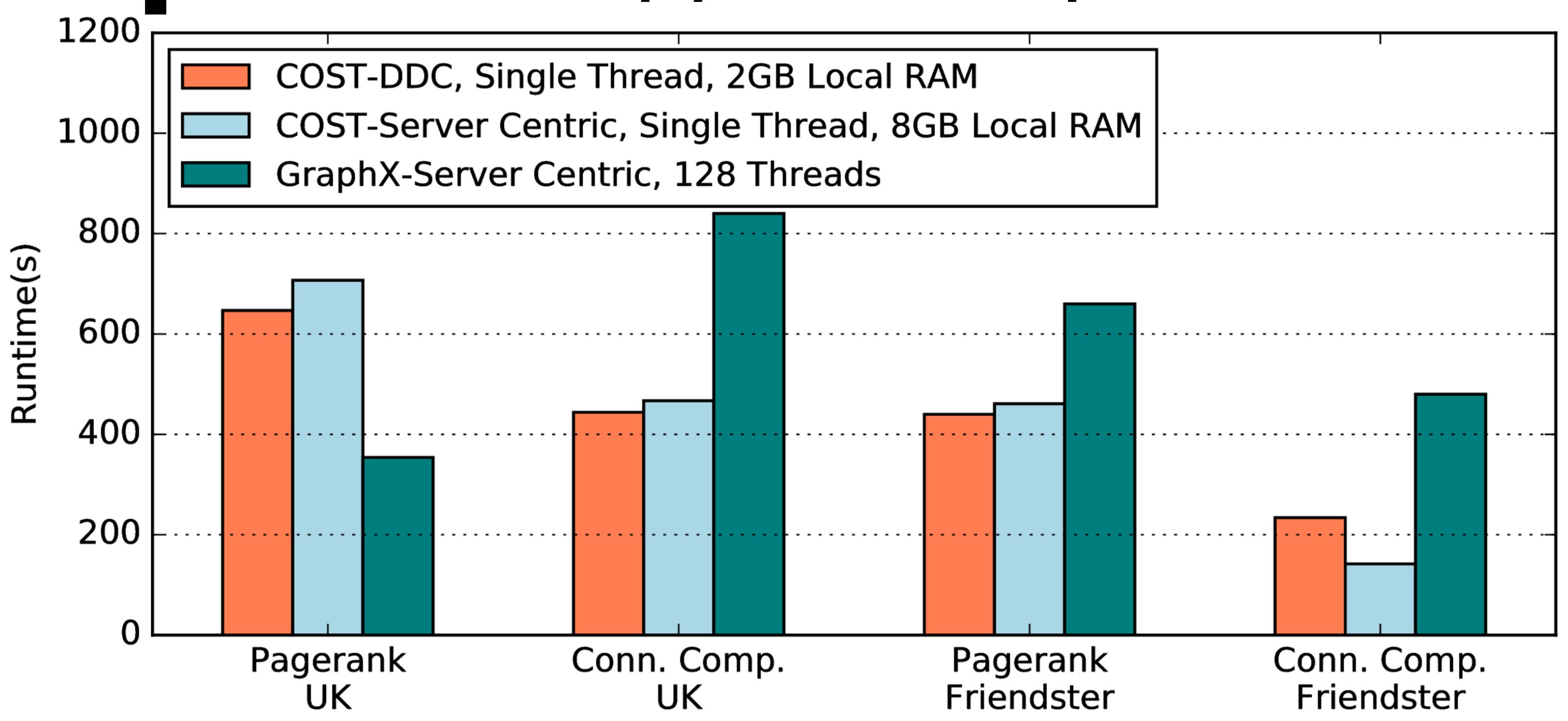


Need efficient transport protocol such as pFabric, pHost



“Closing the loop” by injecting flow FCT to the special swap device. Certain apps require rack scale placement

Improve application performance



Applications can benefit from disaggregation by avoiding coordination and serialization overhead