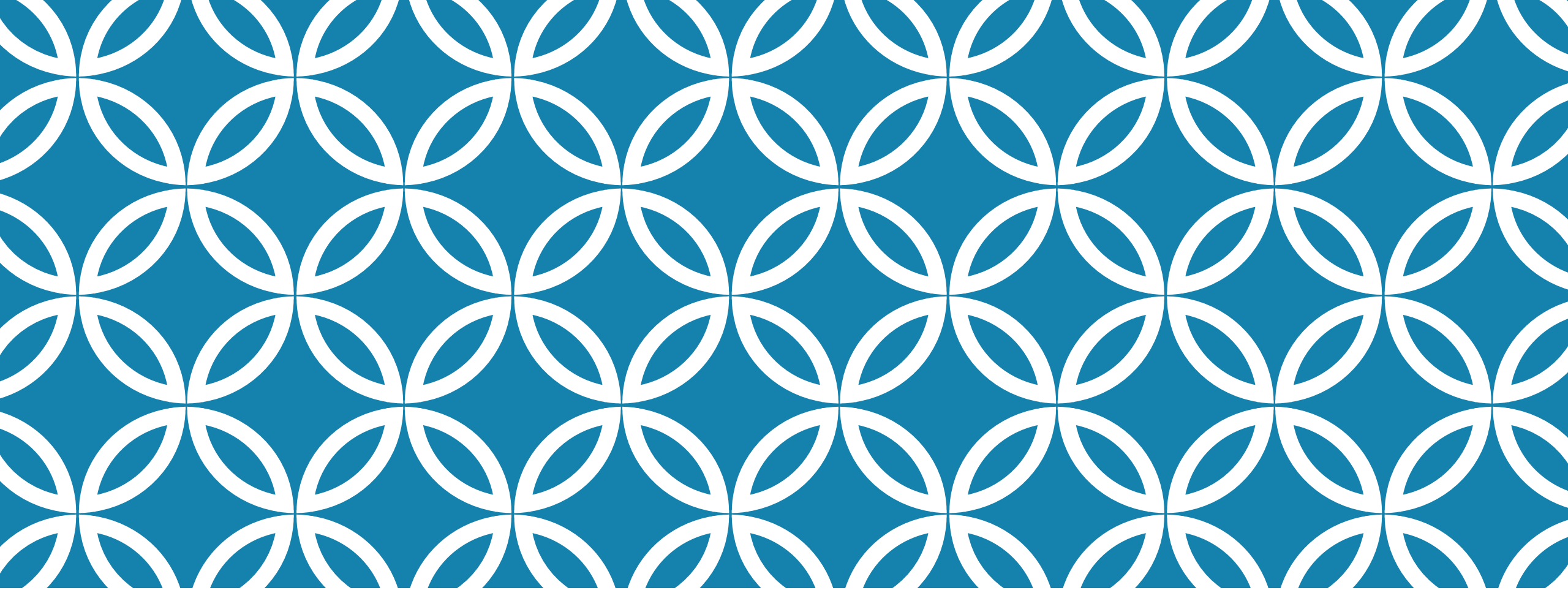


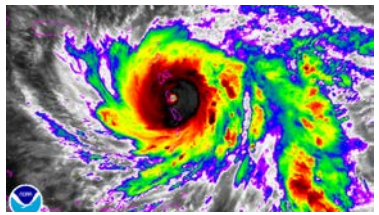
ENABLING CLOUD-HOSTED INTELLIGENCE FOR REAL-TIME IOT APPLICATIONS

Professor Ken Birman
Cornell University
Dept. of Computer Science

DERECHO IS A PLATFORM TO ENABLE MACHINE INTELLIGENCE FOR THE "INTERNET OF THINGS"

IoT devices simply don't have enough power and lack the big picture.

Use the cloud-edge could host machine intelligence, enabling real-time reactivity using consistent, recently-acquired context.



AN EVOLVING CLOUD

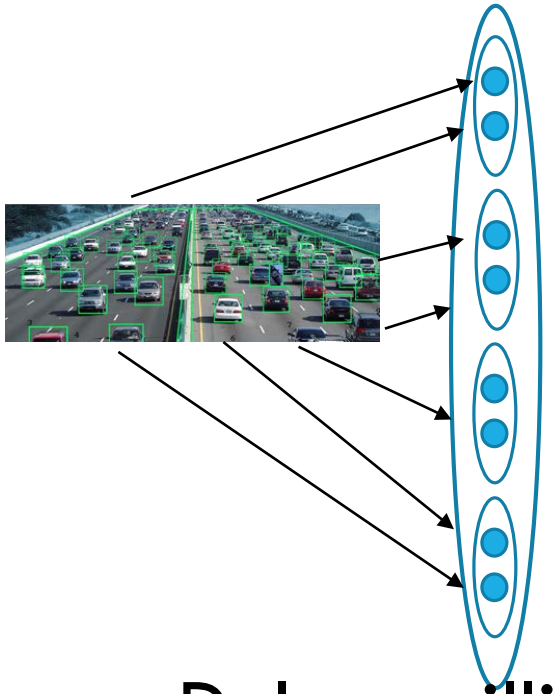
Early generation of cloud solutions: Web pages, advertising

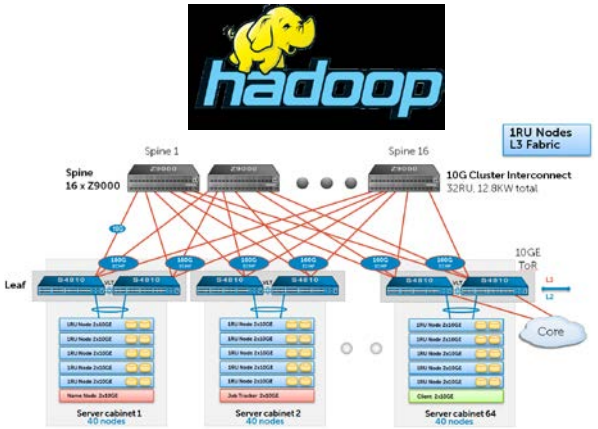
New generation: mobile intelligence, vision, speech understanding

Question to ask: ***Does today's cloud infrastructure fit new demand?***

TODAY: A VERY "LONG" PIPELINE

Data acquisition…. Global File System… Hadoop jobs





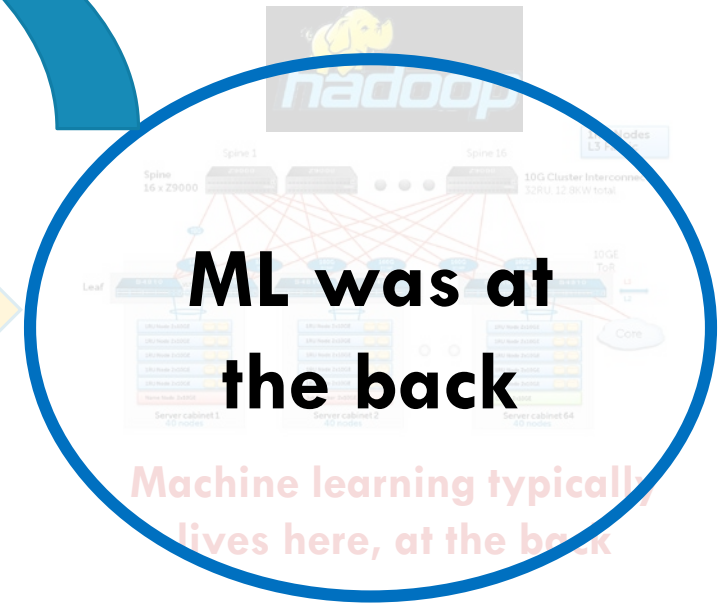
Machine learning typically lives here, at the back

Delay: milliseconds… Seconds…. Hours

NEW: MOVE ML TO THE EDGE OF THE CLOUD

Data acquisition…. Global File Systems Hadoop jobs

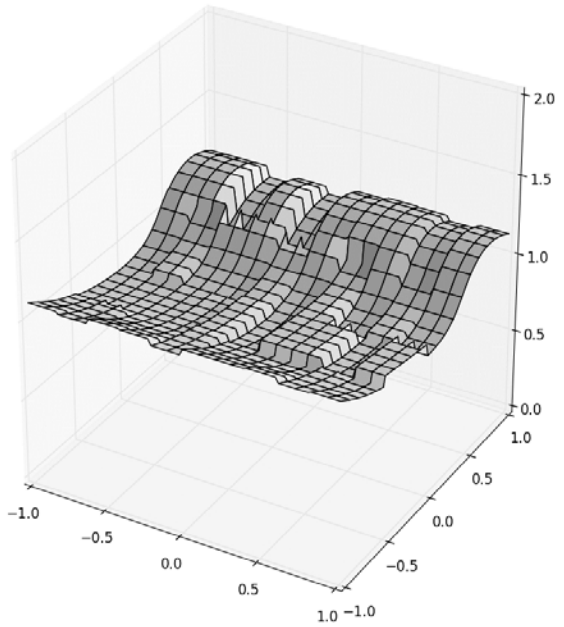
We move data classification and some aspects of learning here



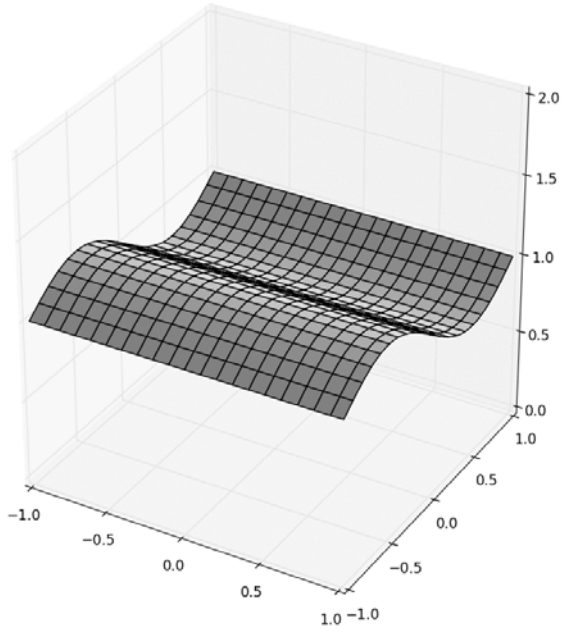
Delay: milliseconds… Seconds…. Hours

THE EDGE REQUIRES SPEED + CONSISTENCY

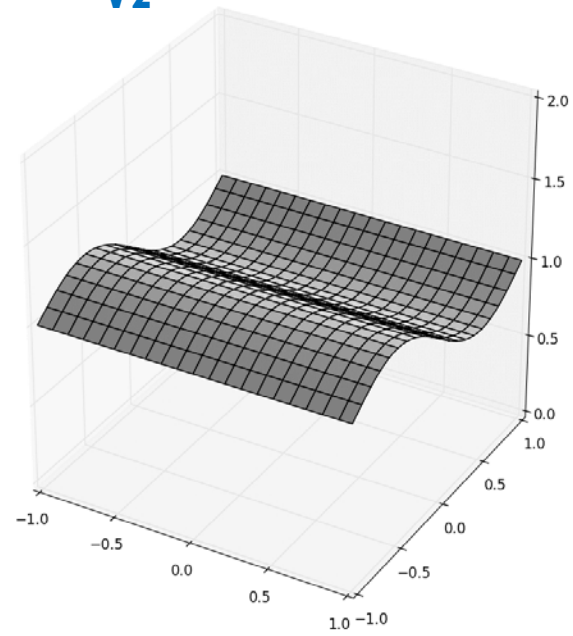
HDFS



$FFFS_{V2}$+SERVER TIME



$FFFS_{V2}$+SENSOR TIME



For real-time IoT data, the Derecho-based storage service ($FFFS_{v2}$) offers optimal temporal accuracy and strong read consistency, lock free.

DERECHO: UNDERLYING PLATFORM

A Derecho

New RDMA software framework for distributed programming (a C++ library)

100x to 10,000x faster than other options

RDMA = "Remote direct memory access".

System

Practice

Theory

It can be used to improve existing cloud μ-services

The first provably optimal Paxos/Atomic Multicast

Zookeeper, HDFS, blob store ($FFFS_{v2}$), BlockChains, DDS…

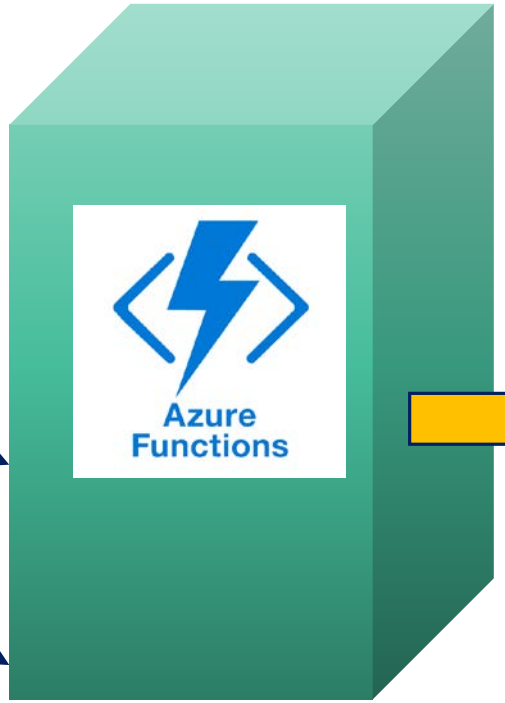
… or your spiffy new smart IoT services

Applying Derecho's asynchronous programming methodology to Paxos led us to the fastest possible protocol

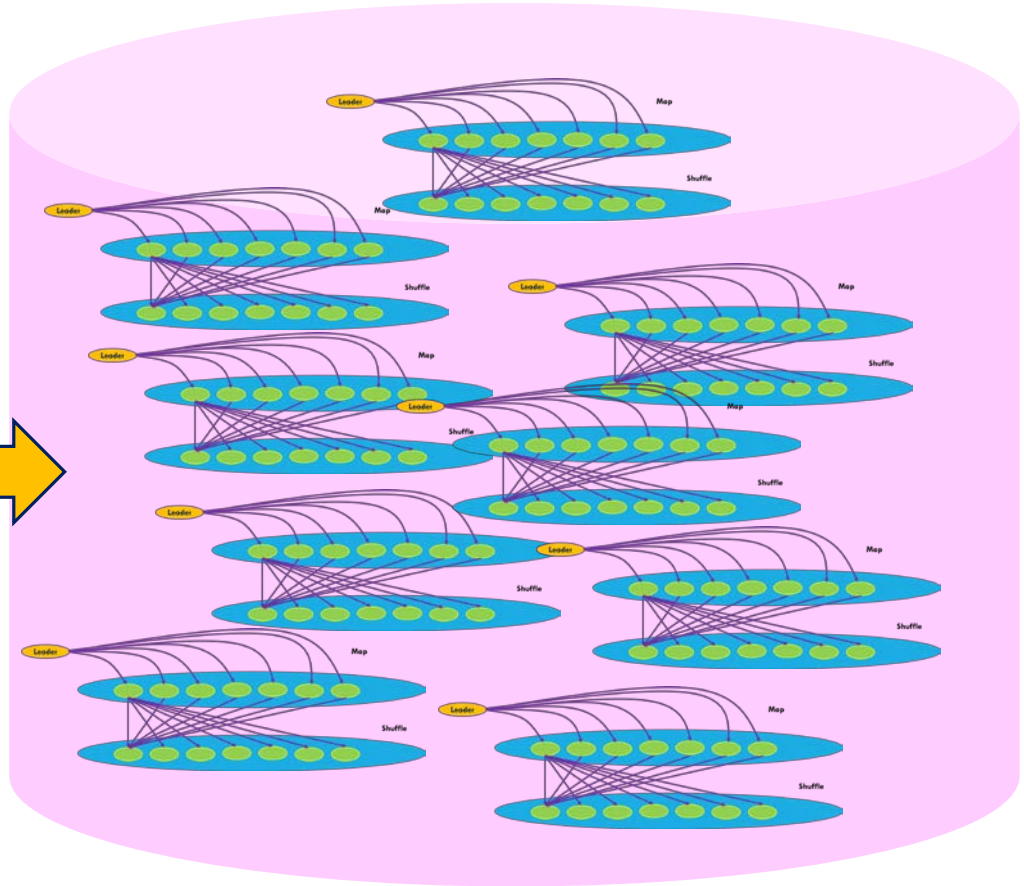
MASSIVELY PARALLEL REAL-TIME USES

REST RPC
Slow but universal

Vast numbers of data sources live outside the cloud itself

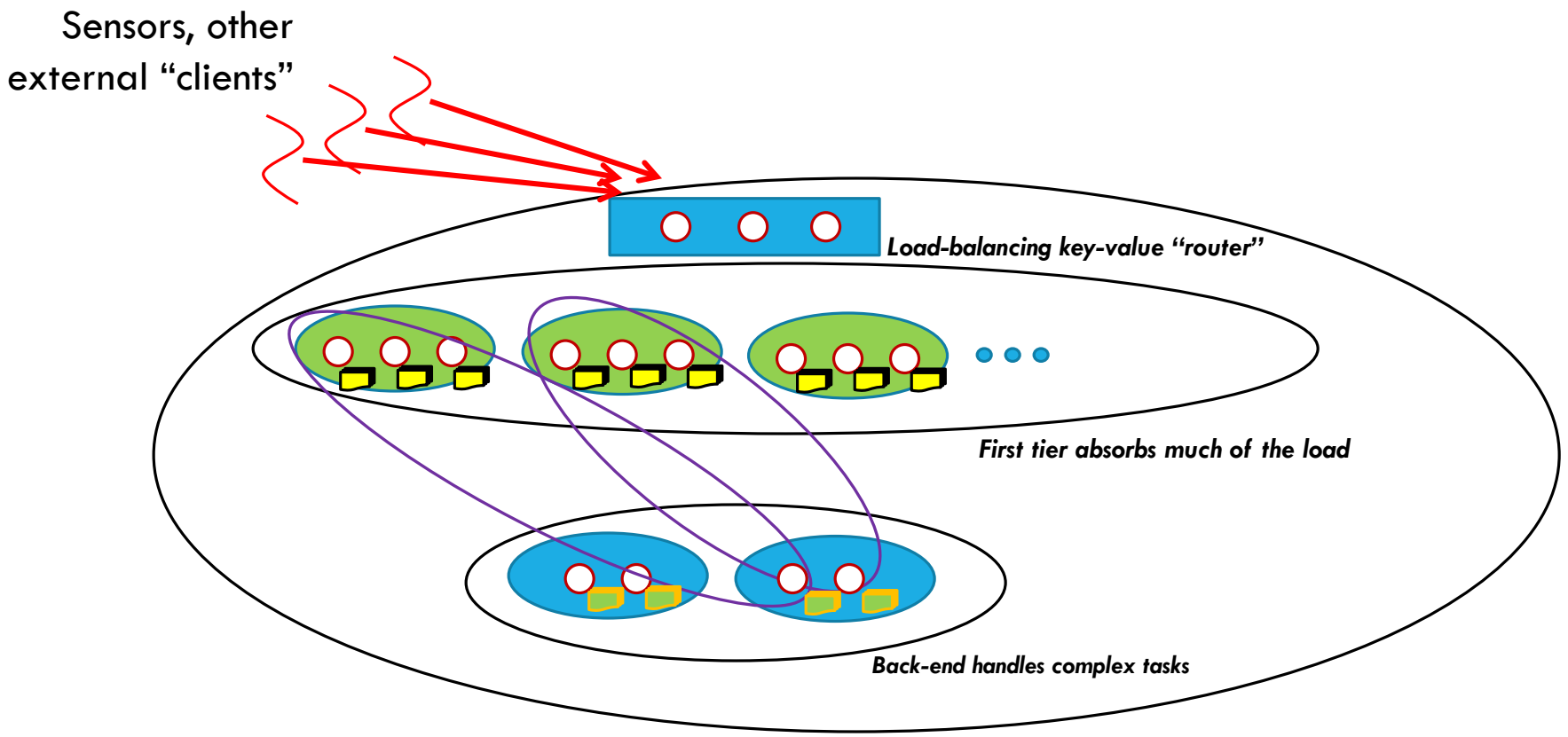


The IoT Cloud uses a tier of lightweight stateless "functions" to absorb load



Derecho is a tool for creating intelligent stateful μ-services, like the Freeze Frame File Server, or this "MapReduce" service

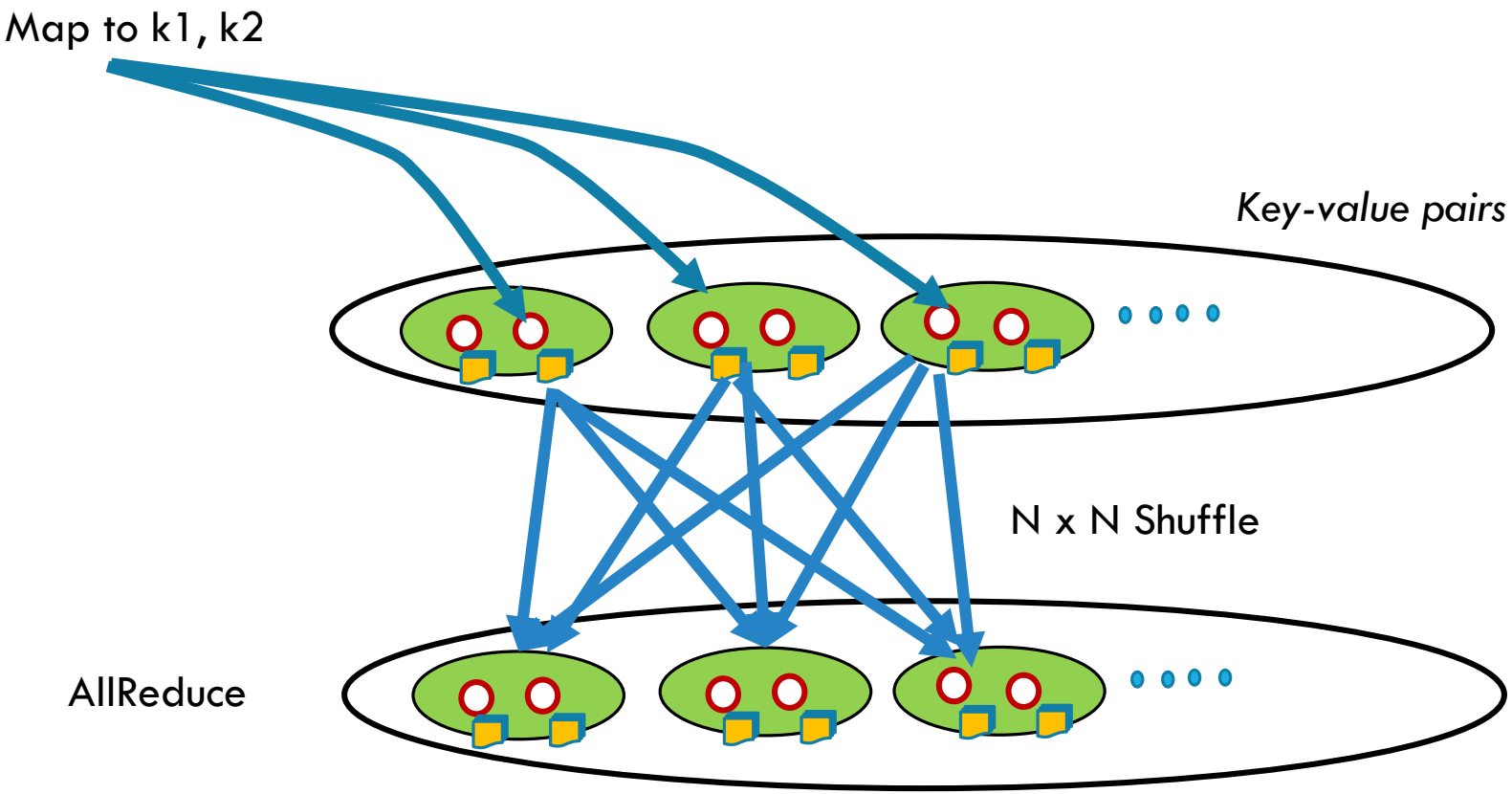
… OUR MODEL: STATE MACHINE REPLICATION IN GROUPS (ATOMIC MULTICAST OR DURABLE LOGGING)



This is just an example.

The developer defines subgroups, controls layout and "shard" pattern

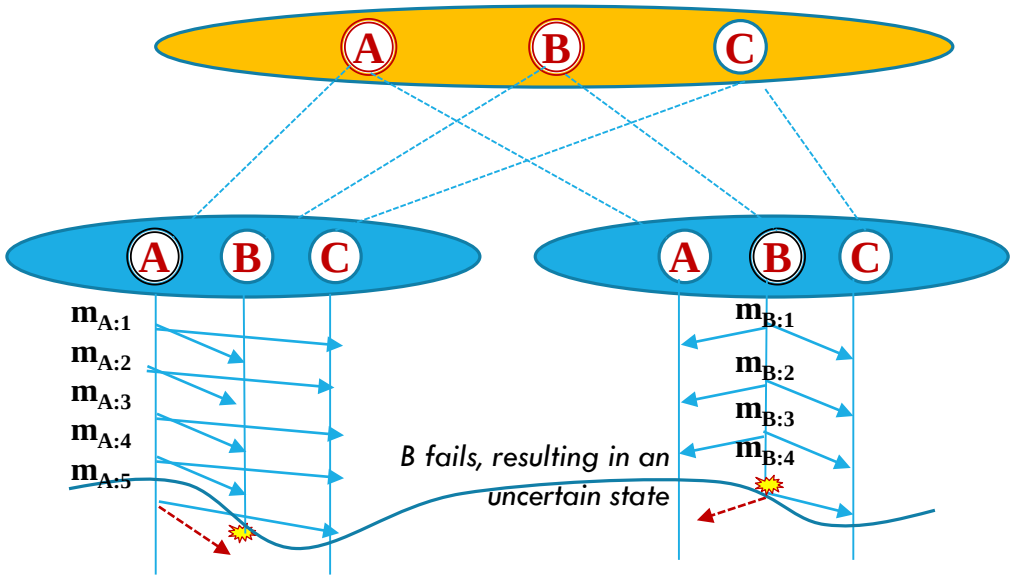
MAP-REDUCE IN A SHARDED GROUP



We obtain a completely atomic MapReduce primitive within Derecho!

IMPLEMENTATION: DERECHO = RDMC/SMC + SST

Derecho group with members {A, B, C}
in which C is receive-only



A, B and C each have a replica of the SST

	Suspected			Proposal	nCommit	Acked	nReceived		Wedged
A	F	T	F	4: -B	3	4	5	3	T
B	F	F	F	3	3	3	4	4	F
C	F	F	F	3	3	3	5	4	F

Data moved on RDMA multicast

Control is done using knowledge programming on the SST

SHARED STATE TABLE: DIRECT RDMA WRITES WITH NO LOCKING (SEQUENTIAL CACHE-LINE CONSISTENCY)

Replicated at members

Update own row

Read-only copy of other rows

	Suspected			Proposal	nCommit	Acked	nReceived			Wedged
A	F	T	F	4: -B	3	4	5	3	0	F
B	F	F	F	3	3	3	4	4	0	F
C	F	F	F	3	3	3	5	4	0	F

RDMA enables A to write directly to the replicas on B and C

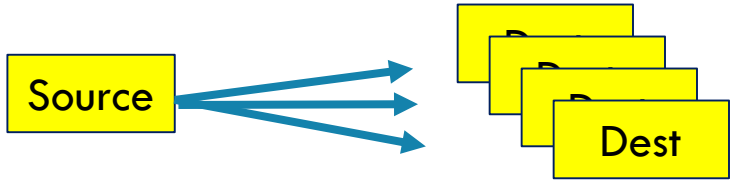
	Suspected			Proposal	nCommit	Acked	nReceived			Wedged
A	F	T	F	4: -B	3	4	5	3	0	T
B	F	F	F	3	3	3	4	4	0	F
C	F	F	F	3	3	3	5	4	0	F

SST PROGRAMMING MODEL

Lock-free, but we store monotonic values in the cells. If you miss some updates you can still deduce that they occurred.

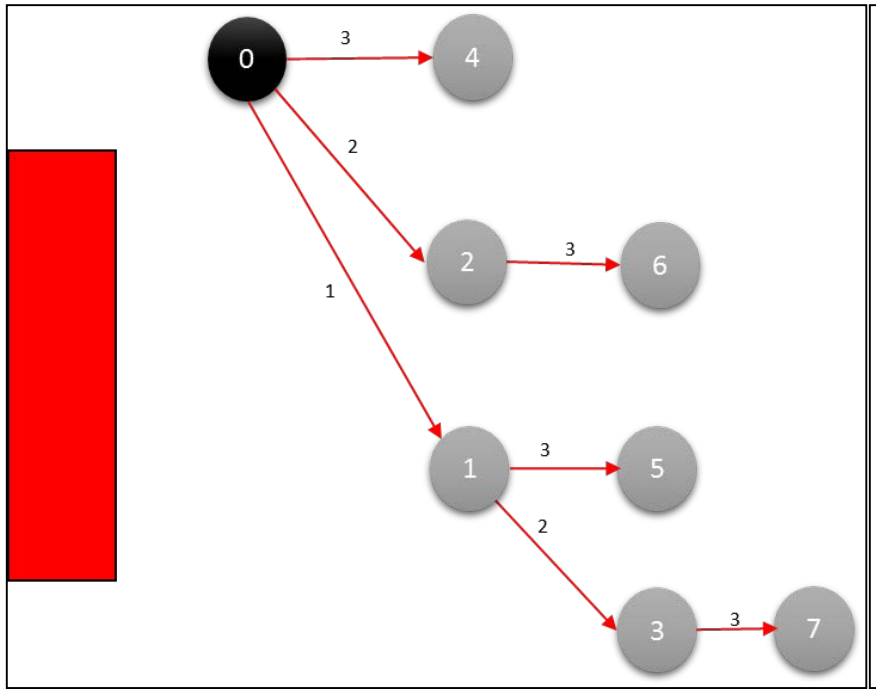
Enables monotonic aggregation and even a monotonic form of knowledge-based reasoning ($\mathrm{K}(\mathcal{P})$, $\mathrm{K}^1(\mathcal{P})$, …).

Result? Highly efficient *batched receiver-side decision-making*.

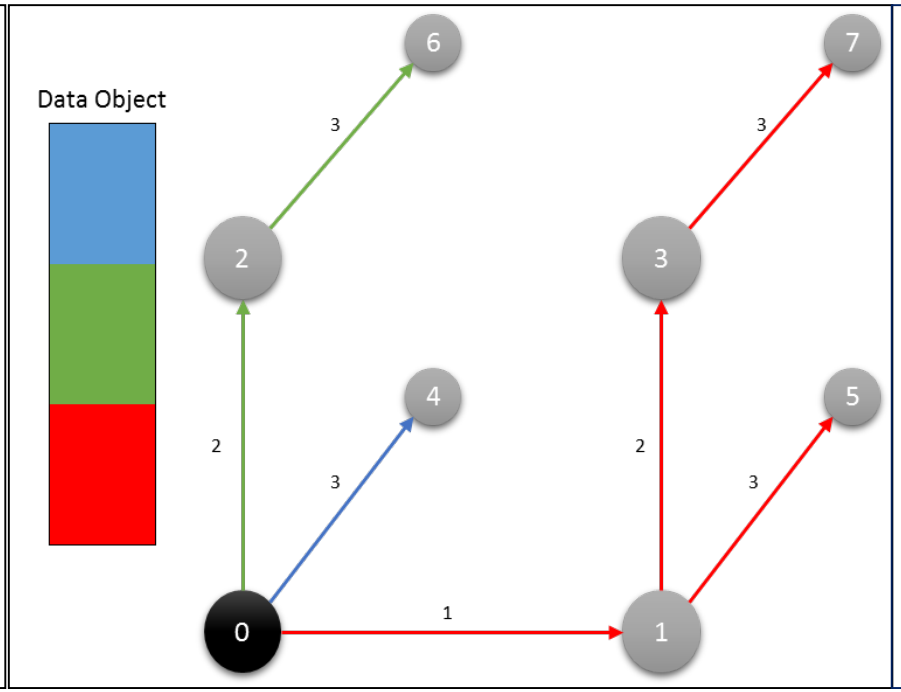


Multicast

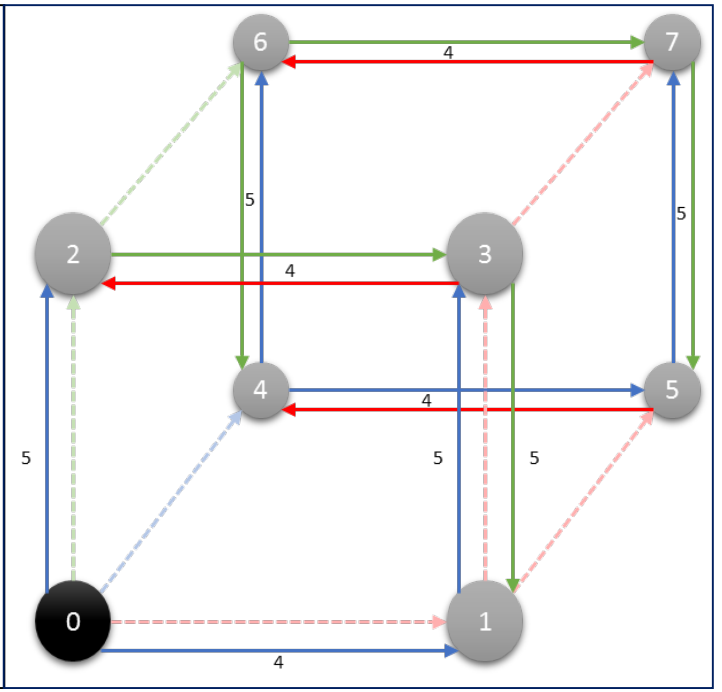
RDMC: AN RDMA MULTICAST



Binomial Tree

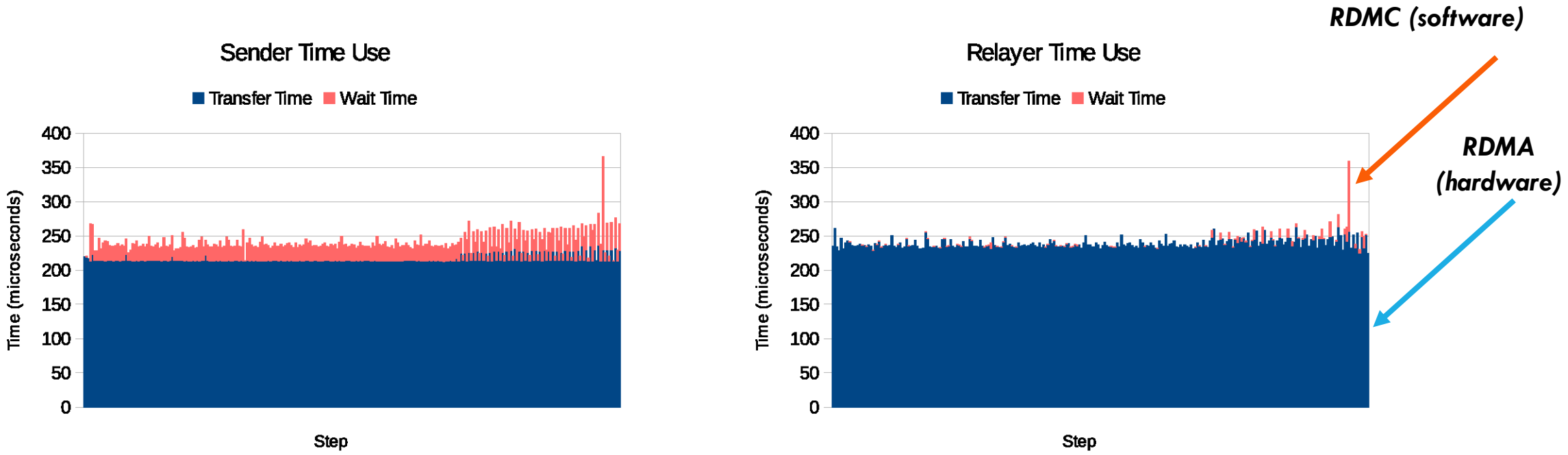


Binomial Pipeline



Final Step

RDMC SUCCEEDS IN OFFLOADING WORK TO HARDWARE



Trace a single multicast through our system… Orange is time "waiting for action by software". Blue is "RDMA data movement".

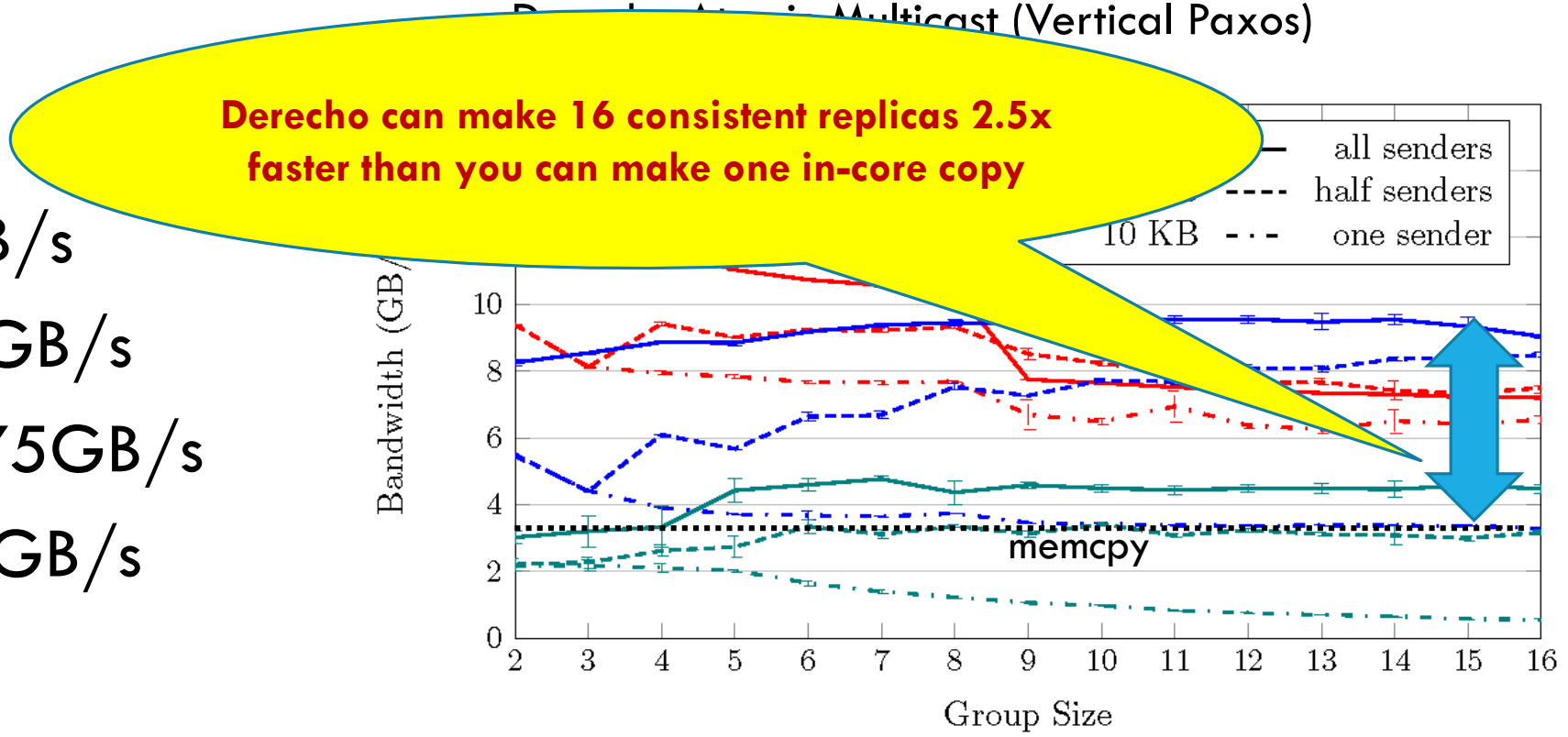
DERECHO IS *FAST*

Mellanox 100Gbps RDMA on ROCE (fast Ethernet)

100Gb/s = 12.5GB/s

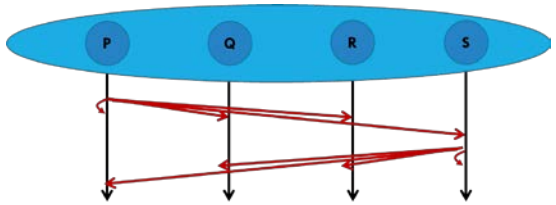
Comparisons:

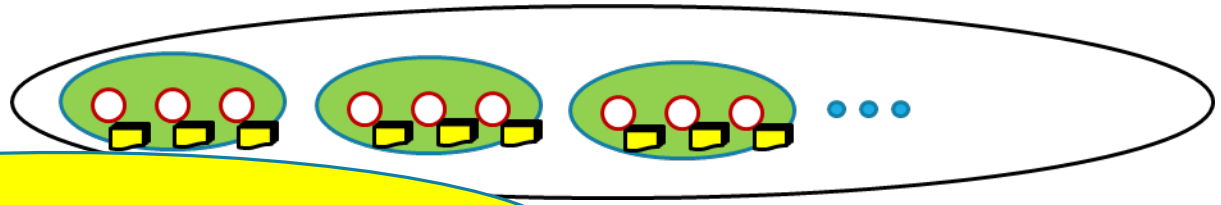
- Derecho: 16GB/s
- Memcpy: 3.75GB/s
- Zookeeper: 0.75GB/s
- LibPaxos: 0.25GB/s



Cool discovery: Derecho outperforms *even on standard TCP.*

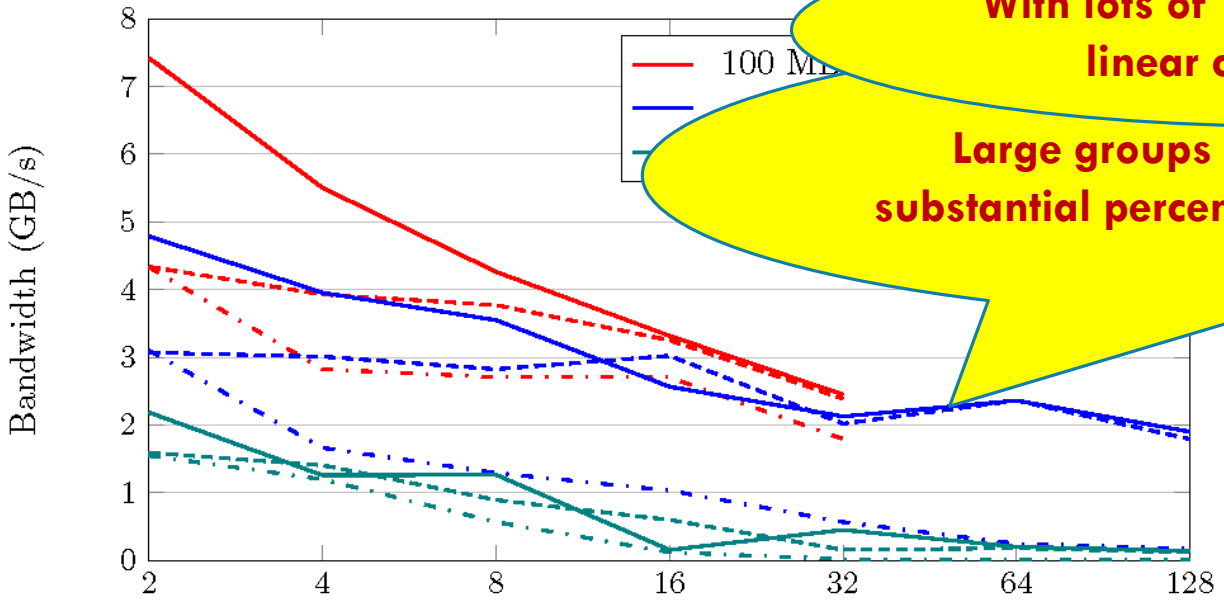
DERECHO: SCALING (56GB/S RDMA)

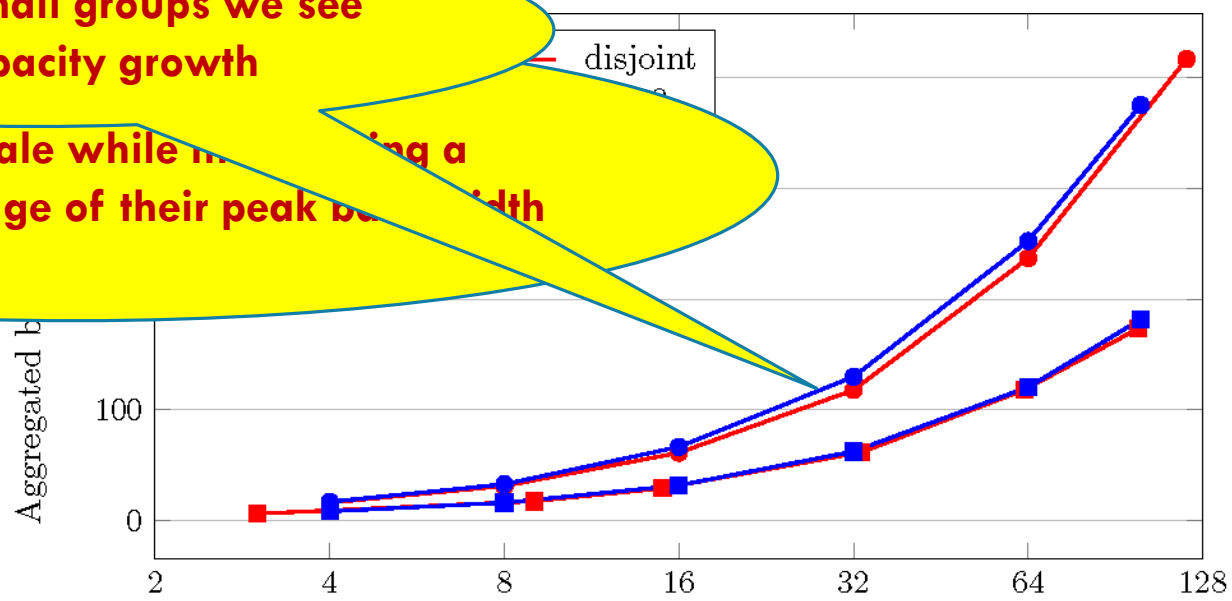




With lots of small groups we see linear capacity growth

Large groups scale while h... ...ng a substantial percentage of their peak b... ...idth





LARGE GROUP OF SIZE N (2...128)
LIMIT WAS MEMORY FOR BUFFERING

BROKEN INTO SHARDS OF SIZE 2 OR 3
LINEAR AGGREGATE THROUGHOUT

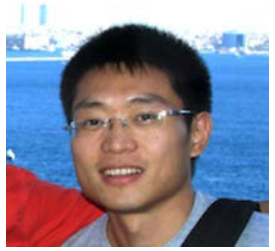
DERECHO PARTICIPANTS

Sagar Jha

Jonathan Behrens*

Matt Milano

Edward Tremel



Weijia Song

Theo Gkountouvas

Ken

Robbert

Derecho: Fast State Machine Replication for Cloud Services. S Jha, J Behrens, T Gkountouvas, M Milano, W Song, E Tremel, S Zink, K Birman, and R van Renesse. 2019. ACM Trans. Comput. Syst. (~March 2019).
RDMC: A Reliable Multicast for Large Objects. J Behrens, S Jha, K Birman, E Tremel. IEEE DSN '18, Luxembourg, June 2018.

* Behrens was a Cornell ugrad, now at MIT pursuing his PhD