some recent trends in distributed systems

Dahlia Malkhi
VMware Research Group (VRG)
VMware Research
founded: DEC 2014

span broad research areas:
- architecture, OS, kernel
- dist. systems, storage, reliability, security
- algorithms, probabilistic analysis, optimization, randomization

research is unfettered and at the same time aims to be aligned with VMware's long term business viability.

publish in top system conferences like SOSP, PODC, NSDI, etc.

David Tennenhouse
Ittai Abraham
Marcos Aguilera
Mahesh Balakrishnan
Dahlia Malkhi
Chris Rossbach
Udi Wieder
Michael Wei (intern, UCSD)
Oracle
DB2
SQL

BigTable
Dynamo
Cassandra

*-DB
dist. caching
dist. transactions

monolithic
rational DBMS

noSQL

newSQL
recent disruptions

- flash and the revival of log-structured stores
- consistent hashing
- memory getting cheaper/larger
- networks getting faster
- CAP theorem
single-node: in-memory map backed by commit-log

purely sequential IO
high-perf random read-access
compaction done post-writing
DHT (Dist. Hash Table):
decentralized, load-balanced data dispersal/retrieval
[Consistent Hashing, Karger et al., 1997]

- localize
- randomize placement
- membership-changes
key-value systems/noSQL
protocol spaghetti

caching, geo-mirroring, versioning, snapshots, rollback, elasticity...
key-value systems/noSQL

BigTable [2006]: irrational tables, weak consistency
Dynamo [2007]: key-value store via DHT, weak consistency
FAWN [2009]: dist-KV backed by SSD, chain replication
Cassandra [2009]: commit after local flush
RAMCloud [2011]: commit before flush (no DHT)
*-DB : …
the CorfuDB shared log design [2011]

CORGU API:
$O = \text{append}(V)$
$V = \text{read}(O)$
$\text{trim}(O) // \text{GC}$
$O = \text{check}() // \text{tail}$

~500K tokens/sec
soft-state, no IO
contention manager
not a point of failure

~500K tokens/sec
soft-state, no IO
contention manager
not a point of failure
Paxos is (in)efficient

Fast-Paxos [L 2006]
Mencius [MJM 2008]
Egalitarian-Paxos [MAK 2013]

Cheap-Paxos [LM 2004]
Vertical Paxos [LMZ 2009]
1-Paxos [DGY 2014]

4 latencies

2 + (F+1) + (F+1) x (F+1) msgs
Paxos leader election is anomalous.

leader-election/membership-change done right:

Virtually Synchronous Paxos
[Lamport, M, Zhou. MSR-TR 2008]

ZooKeeper: Wait-free Coordination for Internet-Scale Coordination
[Hunt, Konar, Junqueira. Reed, Usenix ATC 2010]

Virtually Synchronous Methodology for Dynamic Service Replication

Dynamic Reconfiguration of Primary/Backup Clusters
[Shraer, Reed, M, Junqueira. Usenix ATC 2012]

RAFT: In Search of an Understable Consensus Algorithm
[Ongaro, Ousterhout. Usenix ATC 2014]
• working on multiple objects, wide-area networks, and multi-cores

• Paxos is just too pessimistic: **pre**-determine total order on *everything*..

• ..and this is when Paxos and distributed transactions meet
• txes over totally ordered sequence
  [Percolator 2010, Hyder 2011]

• tx-batches chosen by *mixer* to execute concurrently on multi-cores
  [All about EVE 2012]

• order only among conflicting txes
  [E-Paxos 2013, HyperDex 2012]

• 2-phase-locking with lock-free reads
  [Spanner 2013]

• txes over sequence, distributed protocol helps with resolution
  [CorfuDB 2012]