ZPAXOS: An asynchronous BFT PAXOS with a leaderless synchronous group

D D Amarasekara[1] and D N Ranasinghe[2]

[1]Typefi Systems Pty Ltd, Maroochydore, QLD, Australia
[2]University of Colombo School of Computing, Colombo, Sri Lanka

21ST ISPDC, July 13, 2022, Basel, Switzerland

ZPAXOS: Key features

ZPaxos is a state machine replication protocol which attempts to solve the single leader bottleneck issue and to support BFT with minimum number of replicas

It is a state machine replication protocol based on features of both EPaxos[1] and XPaxos[2]

It achieves consensus in a single round in a non faulty situation with a leaderless synchronous core group

ZPaxos has both improved throughput and latency under both CFT and BFT conditions compared to EPaxos[1]

ZPaxos is based on a single system model to support both CFT and BFT as in XPaxos[2]

[1] Moraru, I., Andersen, D. G. & Kaminsky, M. (2013), There Is More Consensus in Egalitarian Parliaments, SOSP, Farmington, Pennsylvania, USA.
[2] Liu, S., Viotti, P., Cachin, C., Quema, V. & Vukolic, M. (2016), XFT: Practical Fault Tolerance beyond Crashes. Paper presented at the meeting of the OSDI 2016: 12th USENIX Symposium on Operating Systems Design and Implementation, Savannah, GA, USA.

ZPAXOS: Architecture

The replicated state machine comprises a total of **2f +1** replicas;

Out of which the **synchronous core group** consists of **f+1** replicas

Two protocols:

Basic Protocol : PRE_ACCEPT and COMMIT phases

Fault Detection and Recovery Protocol : SUSPECT and ADD_TO_SYNCH_GROUP phases

ZPAXOS: Assumptions

The leaderless core group becomes eventually synchronous

Crash failures are fail-stop only

Every replica of the leaderless synchronous group is eventually correct

There can be at most one faulty CFT or BFT node in the synchronous core group in each protocol round

Only the replicas in the synchronous group accept requests from clients

ZPAXOS: A comparison with EPAXOS[1] and XPAXOS[2]

	CFT	BFT	Replicas	Leader-based	Rounds (No failures)
ZPaxos	✓	✓	2f+1	✗	1
EPaxos	✓	✗	2f+1	✗	2
XPaxos	✓	✓	2f+1	✓	1

Table 1: A comparison for ZPaxos, EPaxos[1] and XPaxos[2]

[1] Moraru, I., Andersen, D. G. & Kaminsky, M. (2013), There Is More Consensus in Egalitarian Parliaments, SOSP, Farmington, Pennsylvania, USA.
[2] Liu, S., Viotti, P., Cachin, C., Quema, V. & Vukolic, M. (2016), XFT: Practical Fault Tolerance beyond Crashes. Paper presented at the meeting of the OSDI 2016: 12th USENIX Symposium on Operating Systems Design and Implementation, Savannah, GA, USA.

ZPAXOS: Other related work

A Leaderless Byzantine Consensus Algorithm[1]:

deterministic, leaderless, partially synchronous, total of 3f+1 replicas

Leaderless Byzantine Paxos[2]:

deterministic, leaderless but with a virtual leader, synchronous, total of 3f+1 replicas

[1] Borran, F. & Schiper, A. (2010), A Leader-Free Byzantine Consensus Algorithm.
[2] Lamport, L. (2011), Leaderless Byzantine Paxos, Springer, DISC 2011: 25th International Symposium, (pp. 141-142).

ZPAXOS: Basic Protocol vs XPAXOS[1]

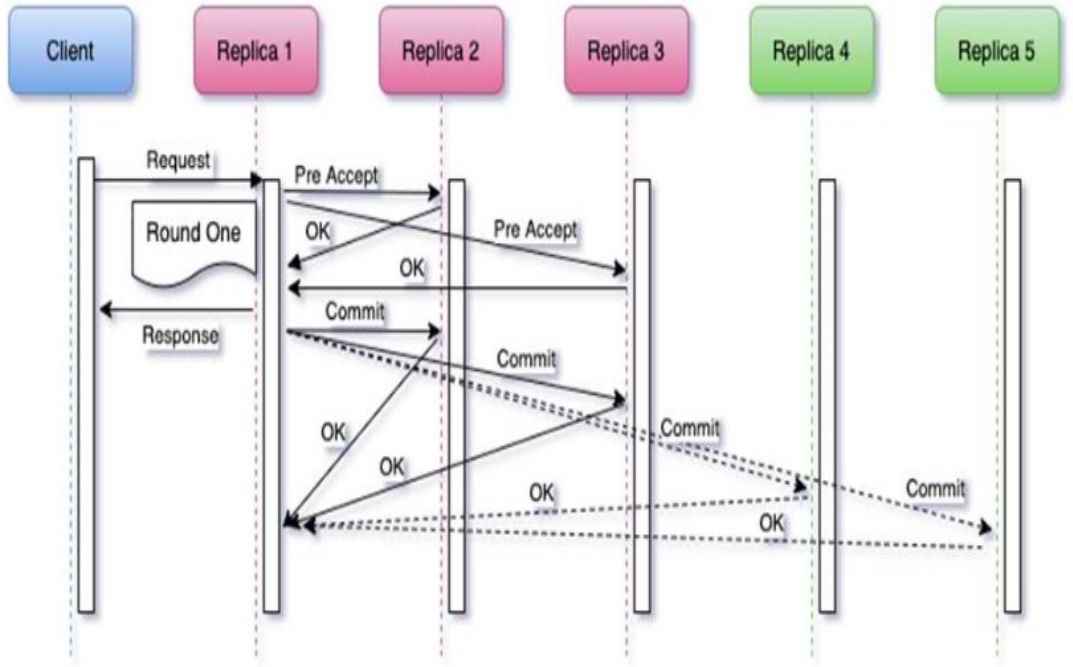


Fig. 1. ZPaxos consensus flow where n=5, f=2, and the synchronous group has three replicas. i.e. 1, 2, and 3.

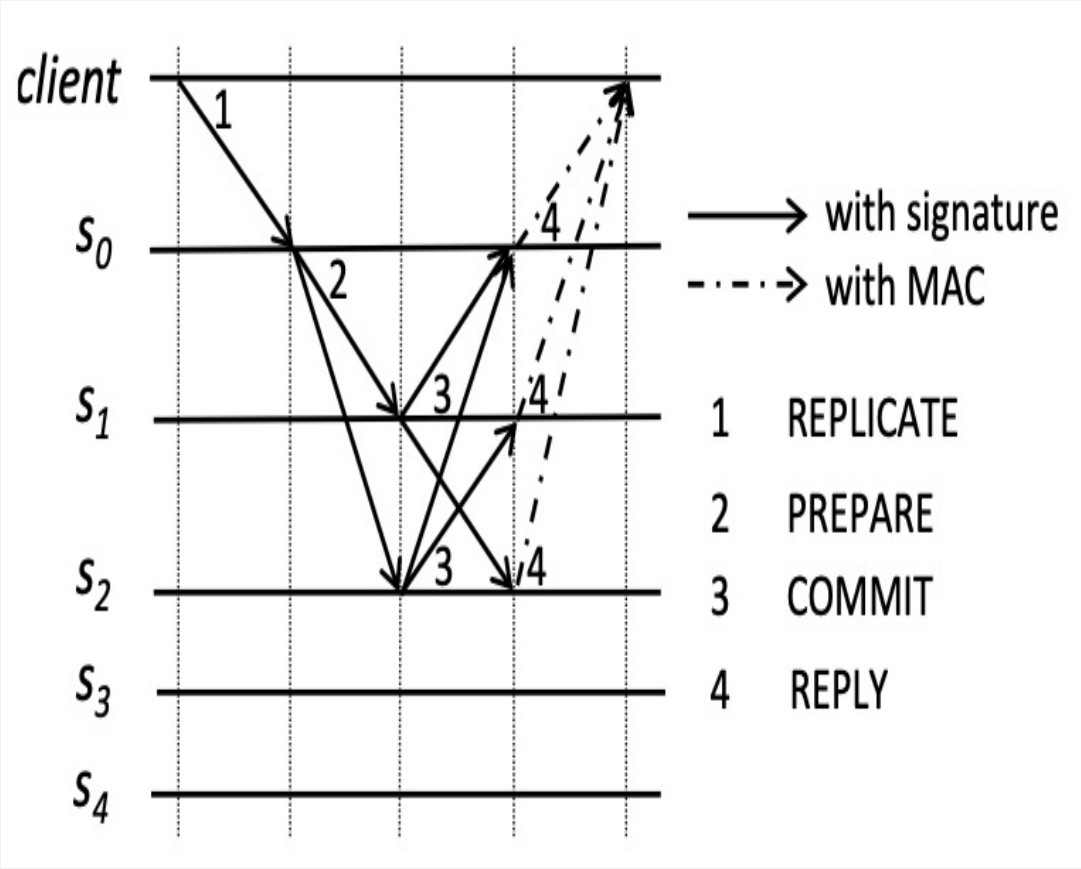


Fig. 1.1 XPaxos Consensus flow where n=5 and f=2

Source: XFT: Practical Fault Tolerance beyond Crashes[1]

[1] Liu, S., Viotti, P., Cachin, C., Quema, V. & Vukolic, M. (2016), XFT: Practical Fault Tolerance beyond Crashes. Paper presented at the meeting of the OSDI 2016: 12th USENIX Symposium on Operating Systems Design and Implementation, Savannah, GA, USA.

ZPAXOS: Basic Protocol vs EPAXOS[1]

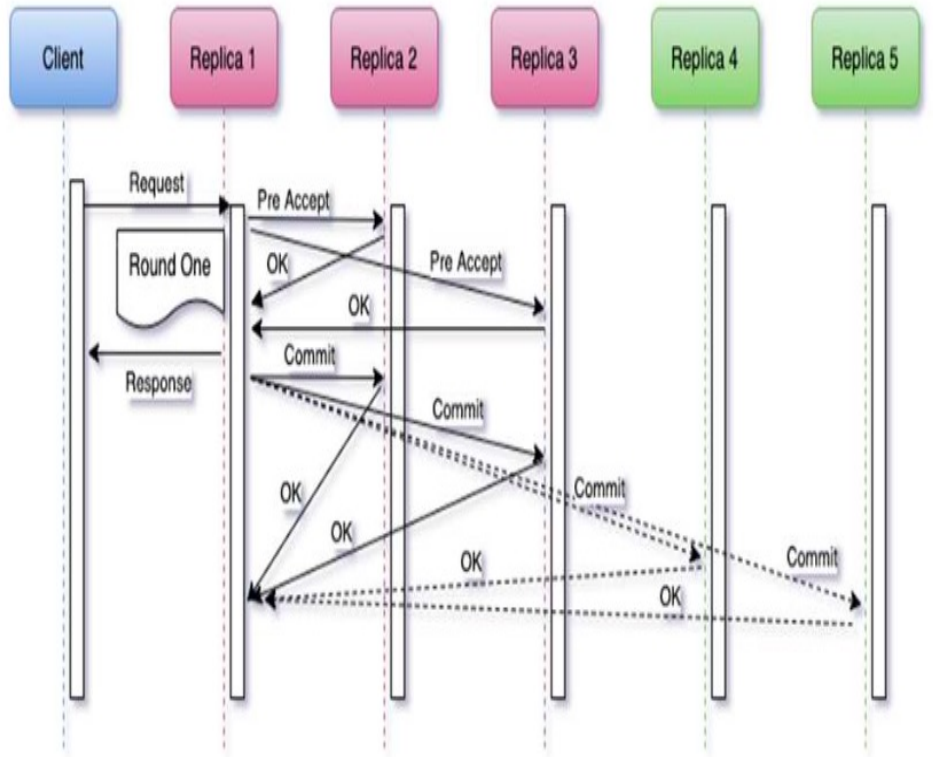
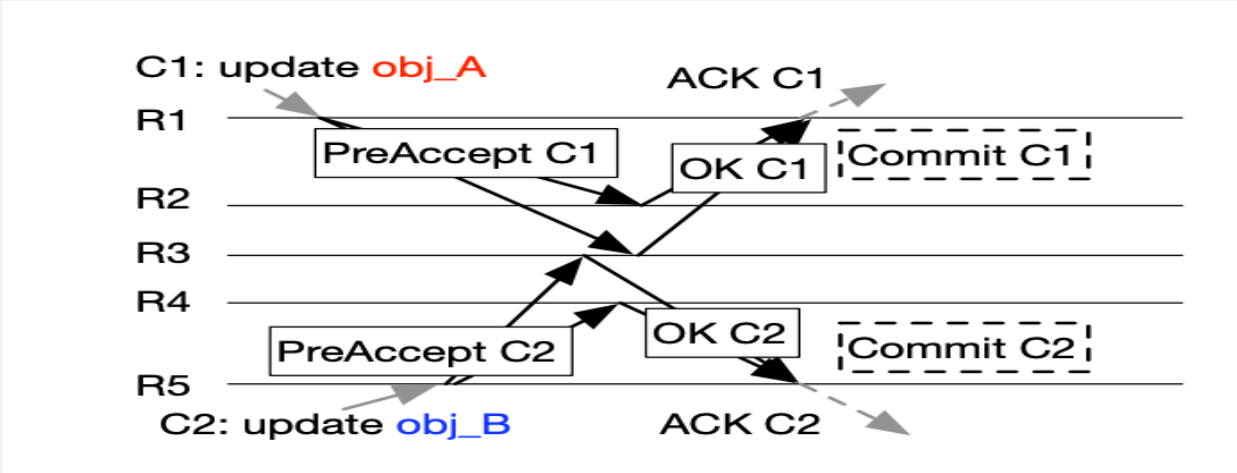


Fig. 1. ZPaxos consensus flow where n=5, f=2, and the synchronous group has three replicas. i.e. 1, 2, and 3.



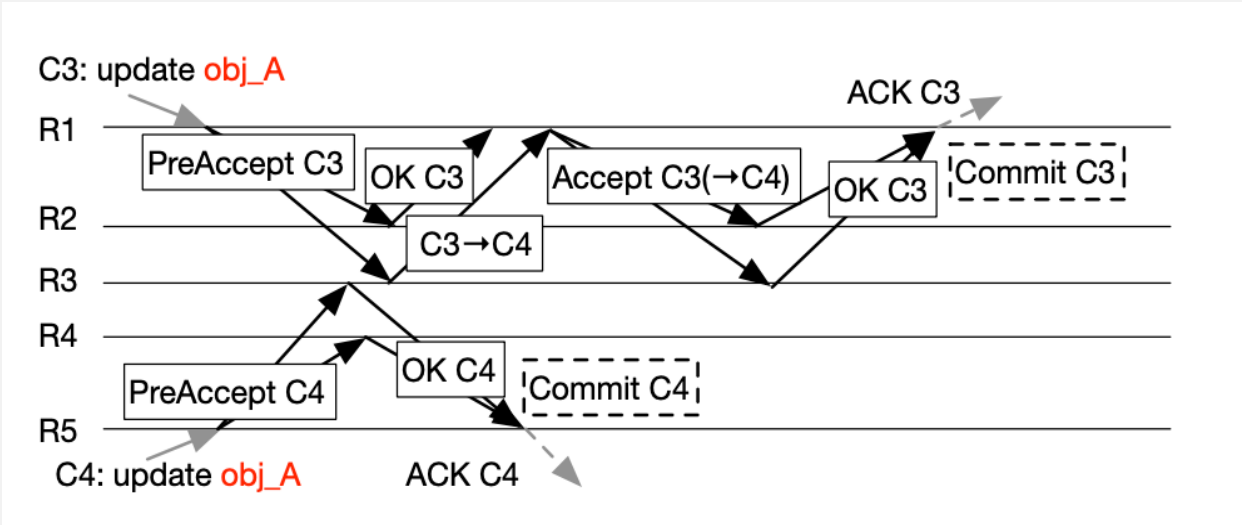


Fig. 1.2 EPaxos Consensus flow where n=5 and f=2

Source: There Is More Consensus in Egalitarian Parliaments[1]

[1] Moraru, I., Andersen, D. G. & Kaminsky, M. (2013), There Is More Consensus in Egalitarian Parliaments, SOSP, Farmington, Pennsylvania, USA.

ZPAXOS: Fault Detection and Recovery Protocol

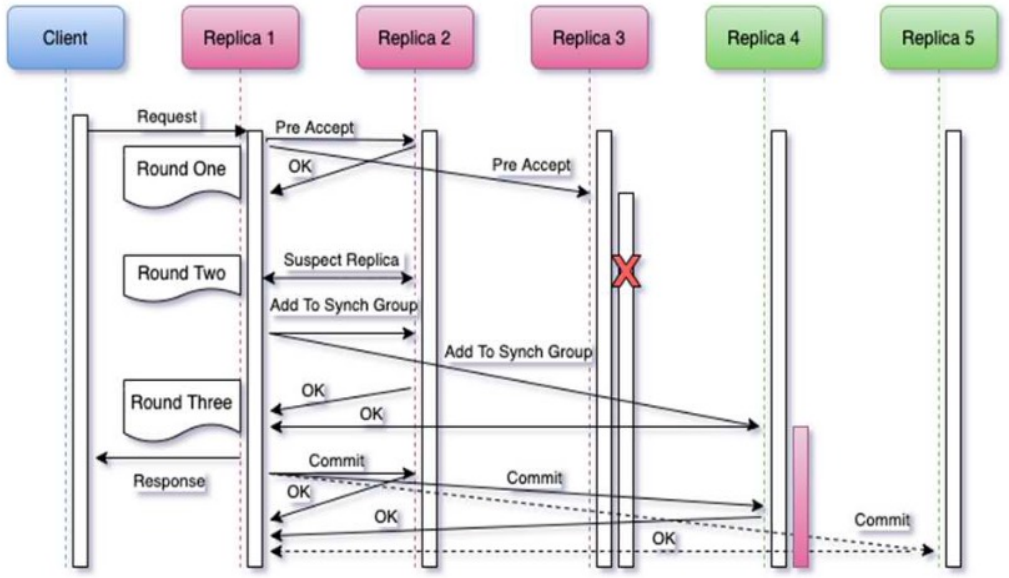


Fig. 2. ZPaxos consensus flow where n=5, f=2, synchronous group size= 3, and a crash failure at replica 3 during the Pre-accept round resulting replica 4 as the new replica to be joined with the synchronous group.

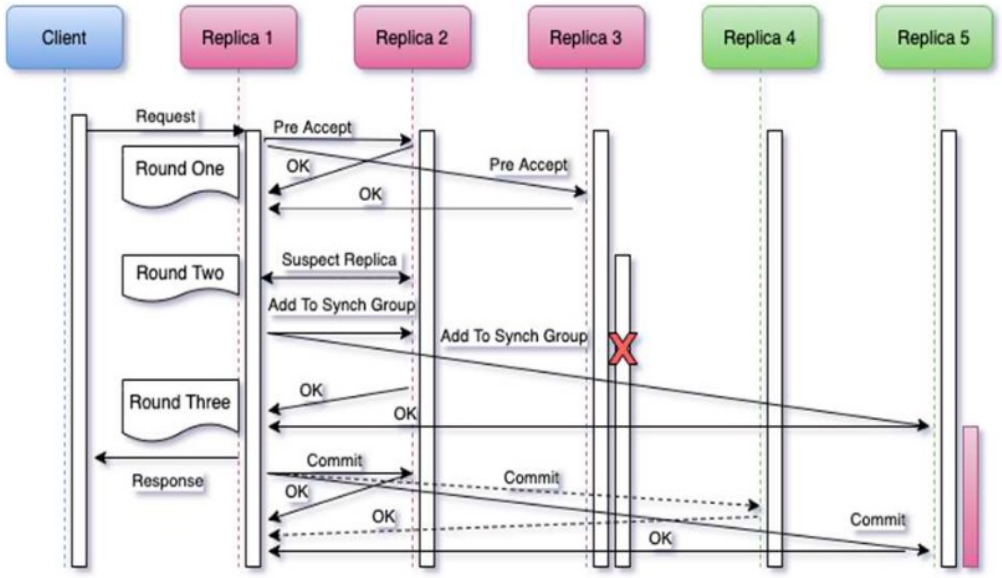


Fig. 3. ZPaxos consensus flow where n=5, f=2, synchronous group size= 3, and a byzantine fault at replica 3 during the Pre accept round resulting replica 5 as the new replica to be joined with the synchronous group.

ZPAXOS: Evaluation

Proposed system was run on a Google Compute Engine service with the below configuration:

A VM instance with machine type e2-medium (2 vCPUs, 4 GB memory), Debian OS; Each replica had own replicated key-value data store

Tests were done on systems having replicas of 3, 5, 7, and 9 for both EPaxos and ZPaxos respectively; The HTTP Benchmark tool **wrk** was used to generate a sufficient load for write operations

Three test cases were considered: no failures, fail-stop failures and byzantine failures

ZPAXOS: Throughput & Latency

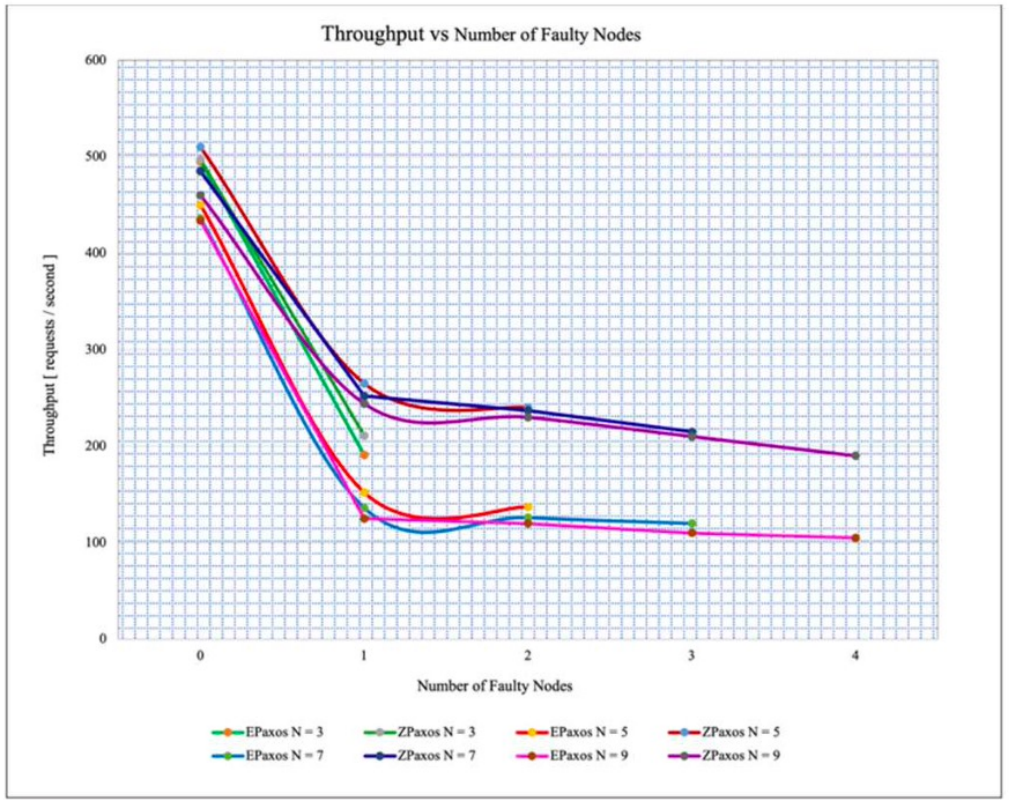


Fig. 4. Throughput vs number of faulty nodes

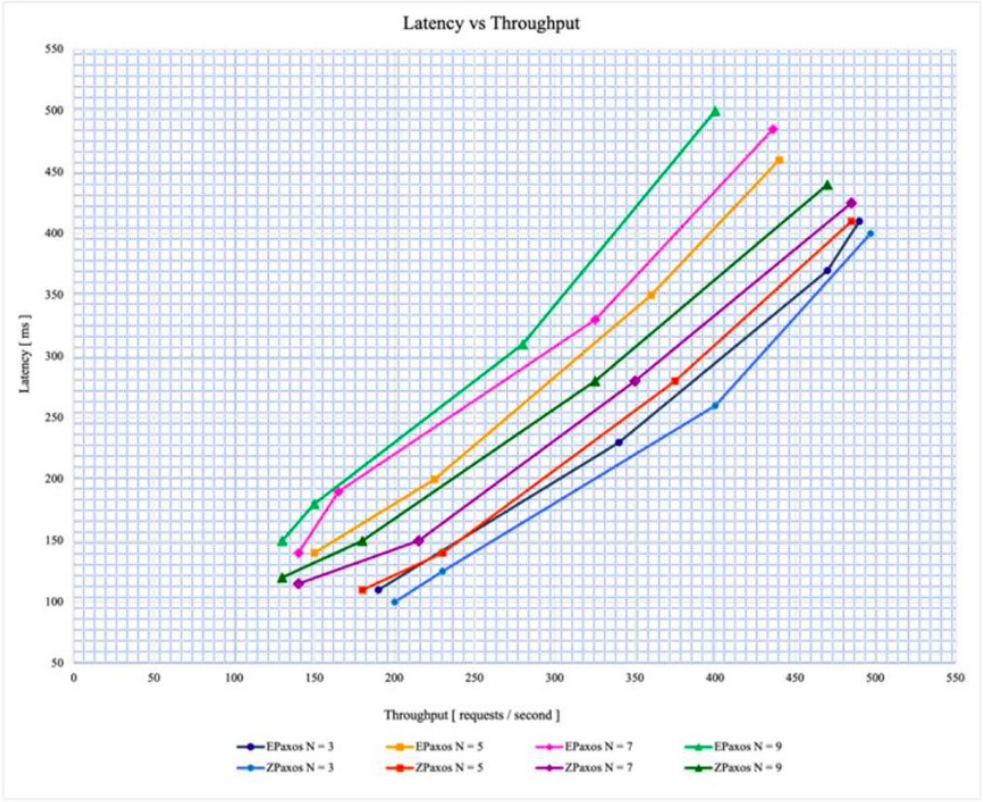


Fig. 5. Latency vs throughput

ZPAXOS: BFT

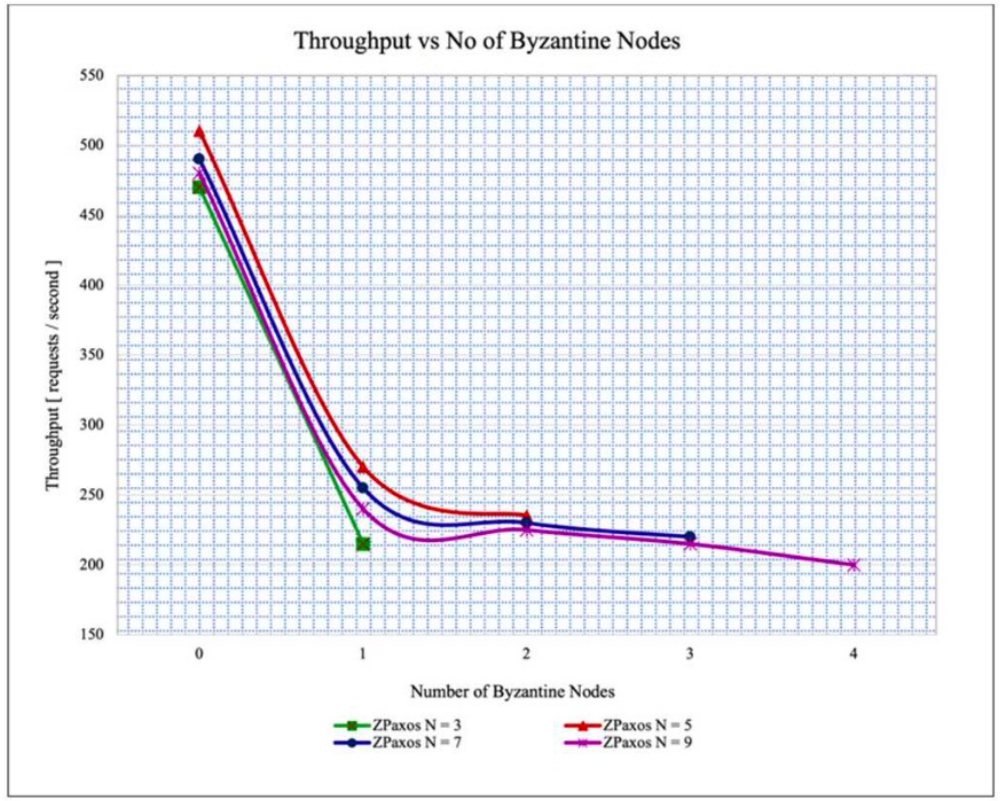


Fig. 6. Throughput vs number of Byzantine nodes

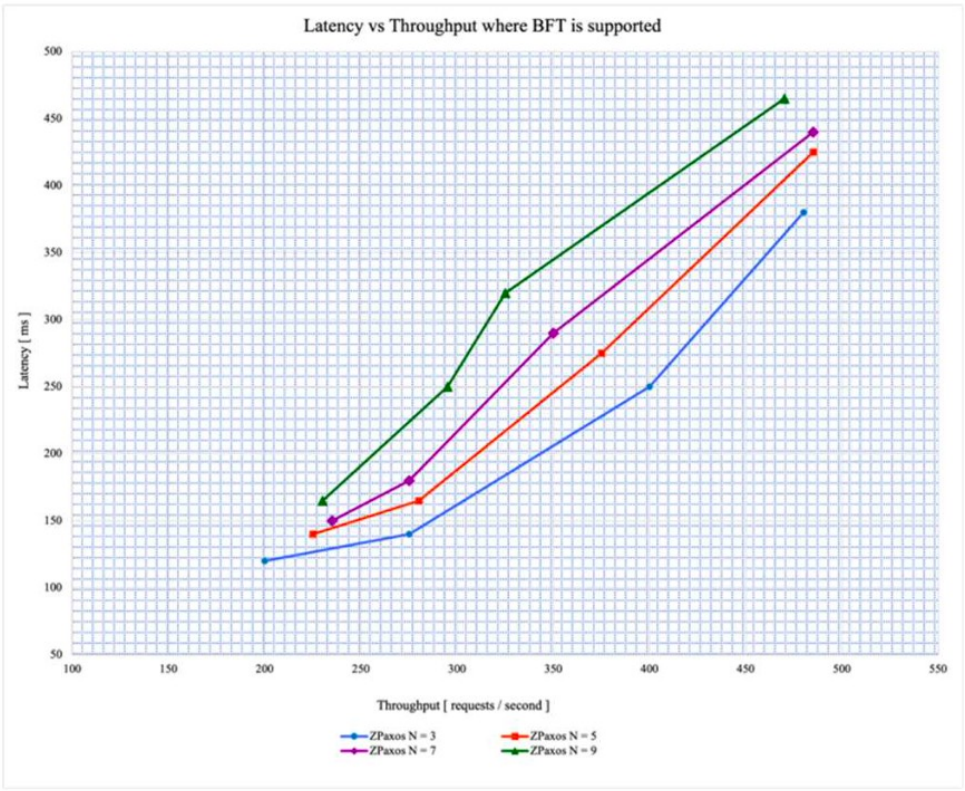


Fig. 7. Latency vs throughput with BFT nodes

ZPAXOS: In the context of PAXOS variants

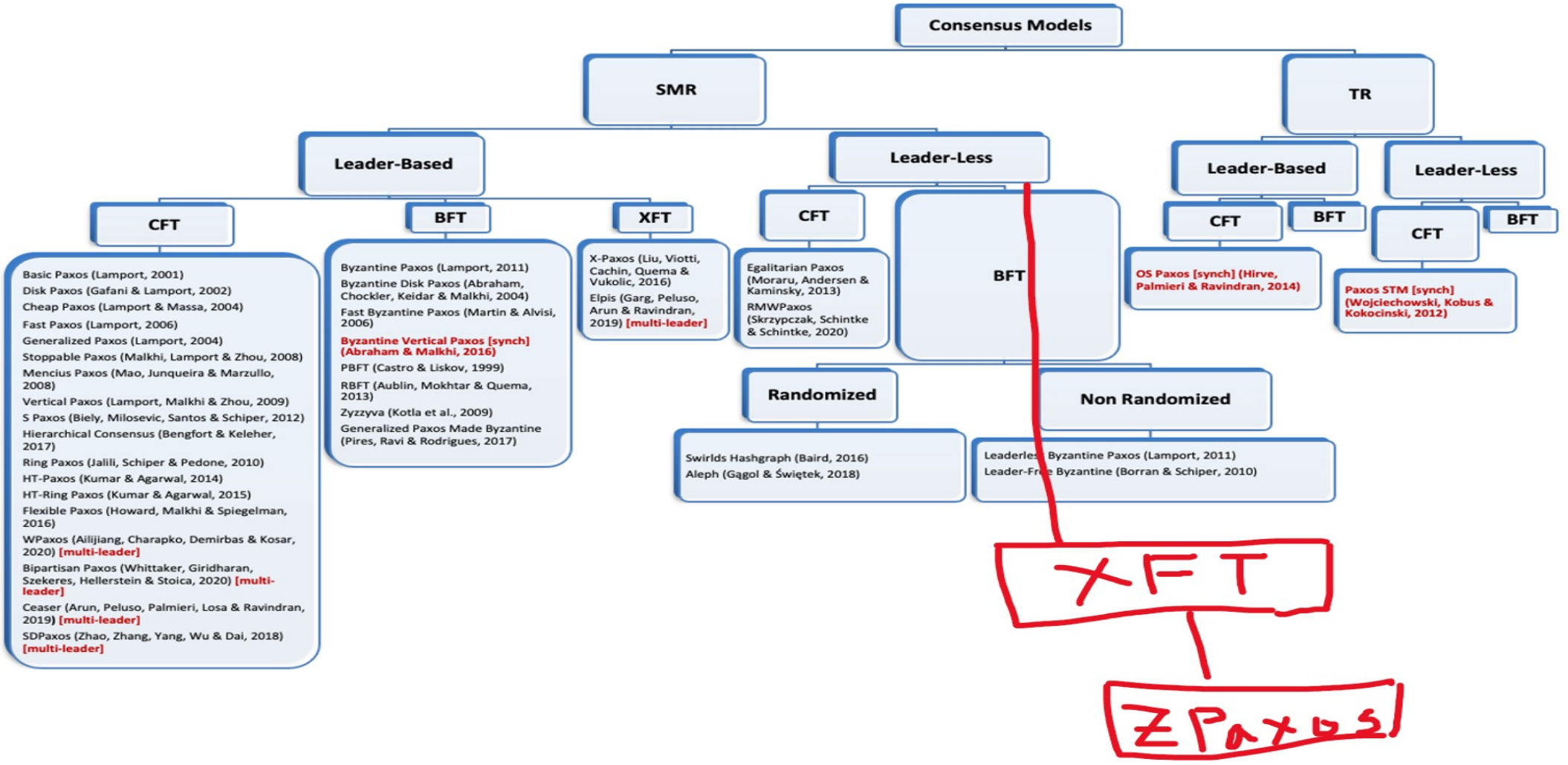


Fig. 8. A Taxonomy of consensus models

ZPAXOS: Future directions

Addressing other failure models such as fail recovery

Handling more than one failure in a consensus round

Can randomization help fast convergence?

ZPAXOS: Q & A

THANK YOU