Distribuované systémy

Štefan Šafár

Obsah

- Proč distribuované systémy
- Sjednocení pojmů
- CAP theorem
- Replikace

- http://book.mixu.net/distsys/single-page.html

Když server nestíhá

- Upgrade HW
- Co pak?

Škálovatelnost

Schopnost systému zvládnout zvětšující se množství práce rozumným způsobem

- Velikostní
- Geografická
- Administrativní

Výkon

- Nízká latence
- Vysoká propustnost
- Nízké využití výpočetního výkonu

Dostupnost

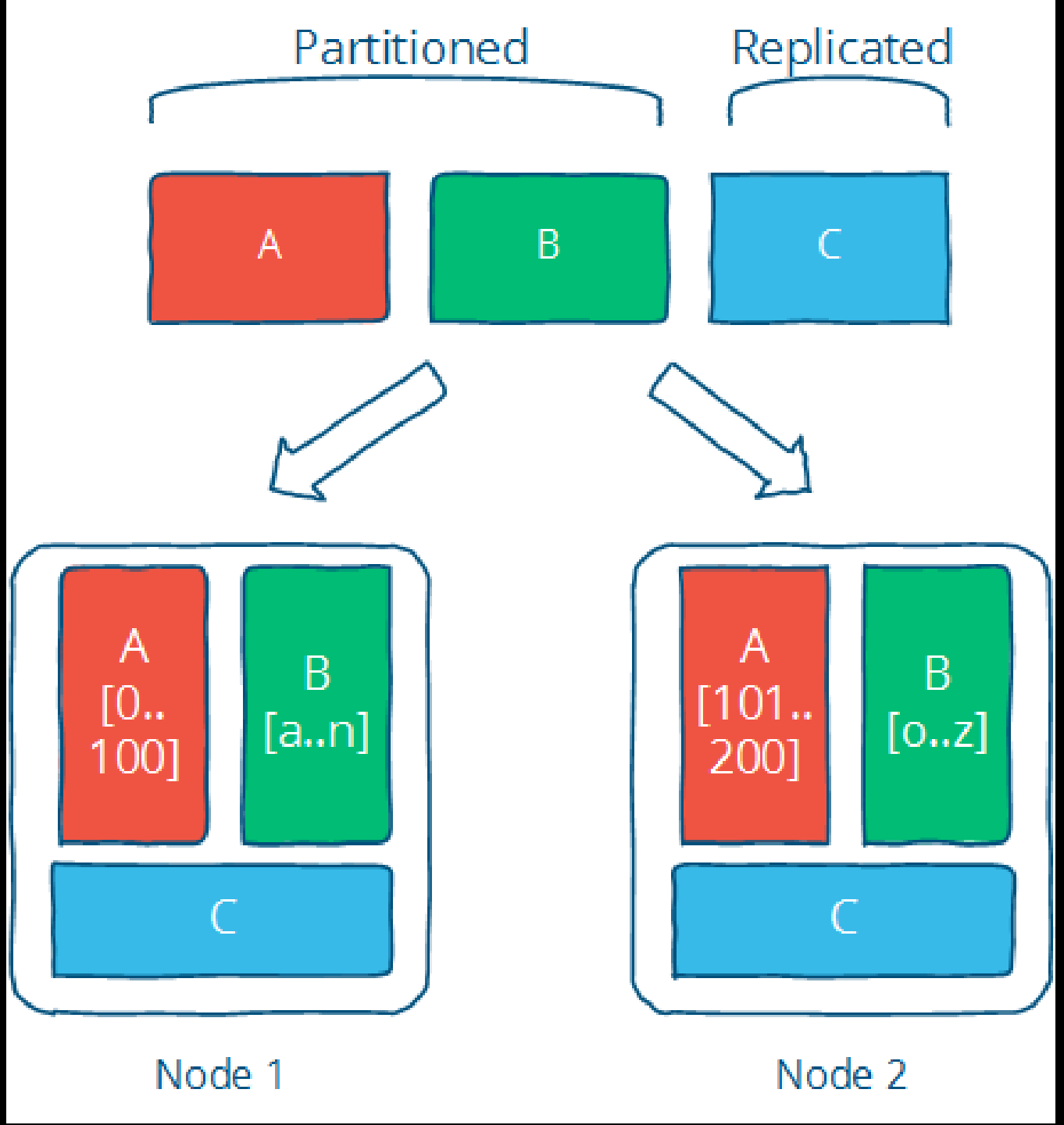
- Procento času, po který byl systém „funkční“
- Jeden stroj nemůže být fault tolerant

Fault tolerance

- Schopnost systému chovat se „nějak“ při výpadku
- Výpadku čeho?

Důsledky distribuovanosti

- Víc uzlů => větší pravděpodobnost selhání
- Víc uzlů => víc komunikace mezi uzly
- Geografická vzdálenost => latence


Partitioned
Replicated
A
B
C
A
[0..
100]
B
[a..n]
C
A
[101..
200]
B
[o..z]
C
Node 1
Node 2

Partitioning

- + výkon (divide & conquer)
- + dostupnost celku (fail independently)

Replikace

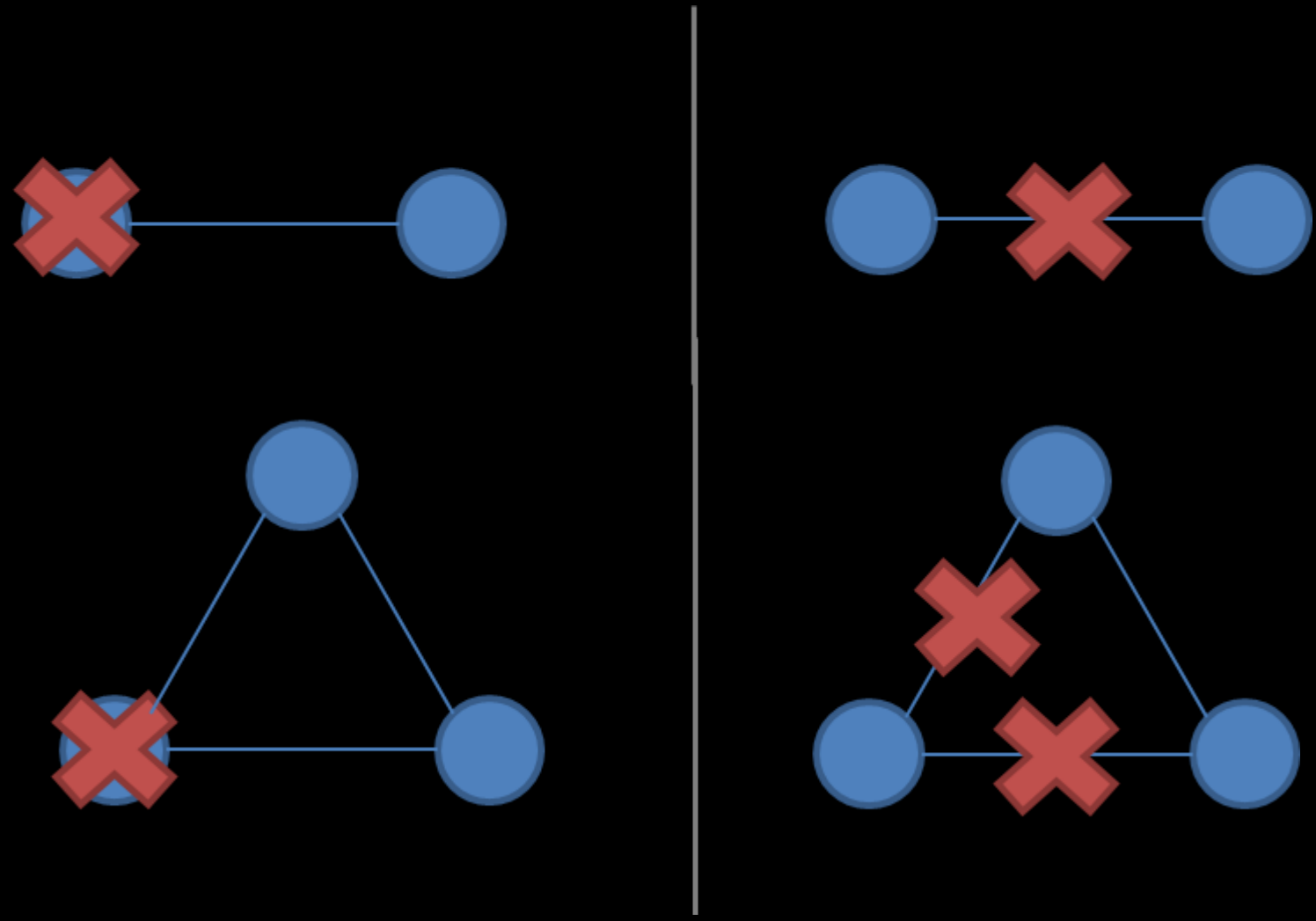
- + výkon (víc nodů může řešit stejný problém)
- + dostupnost (n+1 uzlů)
- konzistence ?

Model systému

Sada předpokladů o systému

- Co vše může selhat
- Pořadí operací
- Lokální vs vzdálený stav
- Synchronizace času / dat
- Pád sítě / pád uzlu

Network partition



Synchronní systém

- Lock, write, unlock
- Omezený výkon

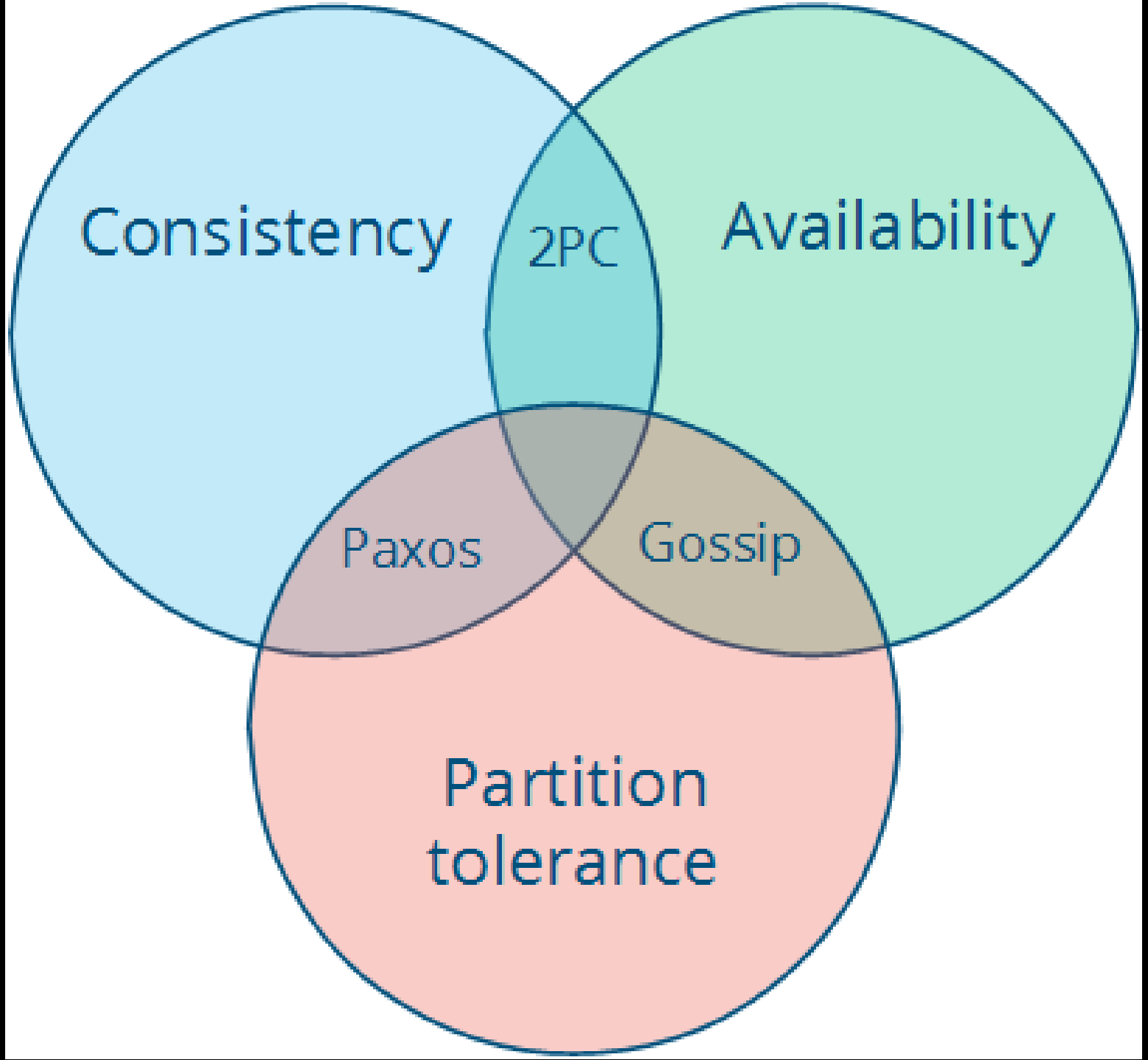
Asynchronní systém

- Žádný předpoklad o časování/pořadí

Problém konsenzu

Uzly dosáhnou konsenzu, pokud se dohodnou na nějaké jedné hodnotě.

- Dohoda
- Integrita
- Ukončení
- Validita

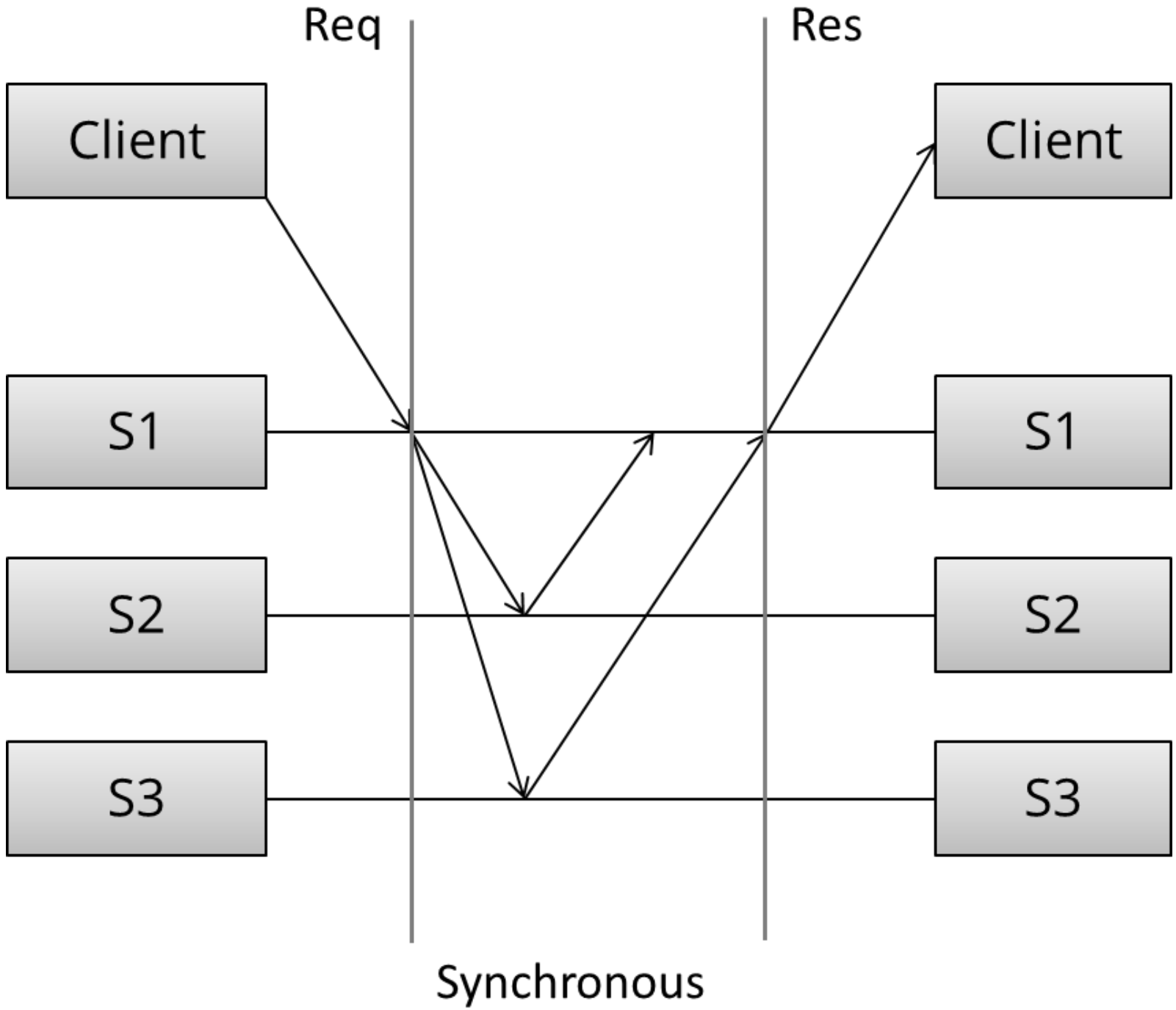

Consistency
2PC
Availability
Paxos
Gossip
Partition
tolerance

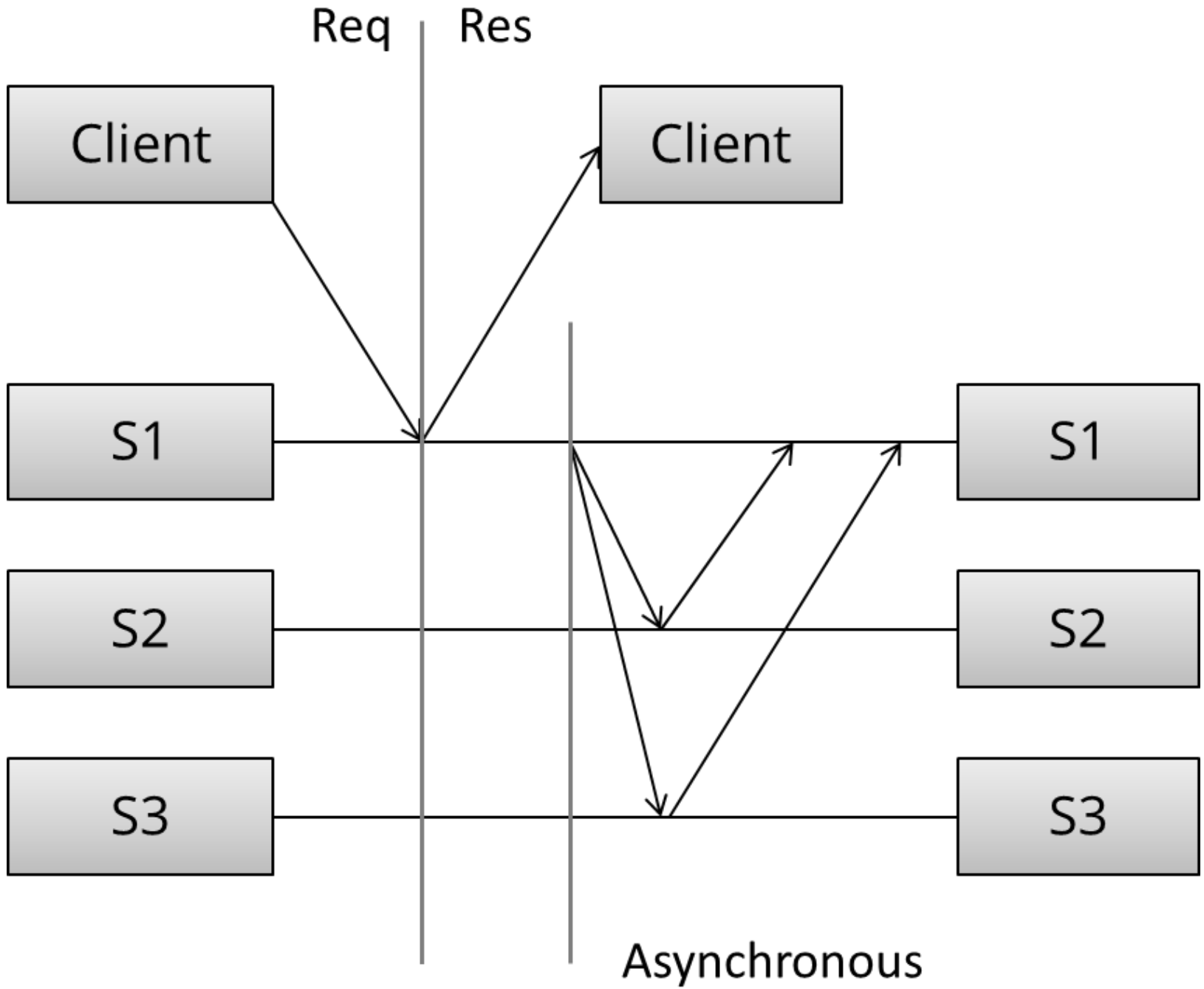
Pozorování

- Mnoho starších databázových systémů je CA
- Musíme si vybrat A nebo C při network partition
- Silná konzistence vs výkon
- Pokud chceme A+P, musíme snížit C

Pořadí a čas

- Úplné seřazení (total order)
- Částečné (partial order)
- Čas => pořadí
- Globální / lokální čas
- Serializovatelnost
- Detekce selhání (failure)


Req
Res
Client
Client
S1
S1
S2
S2
S3
S3
Synchronous


Req
Res
Client
Client
S1
S2
S3
S1
S2
S3
Asynchronous

2-phase commit

[Coordinator] -> OK to commit? [Peers]

<- Yes / No

[Coordinator] -> Commit / Rollback [Peers]

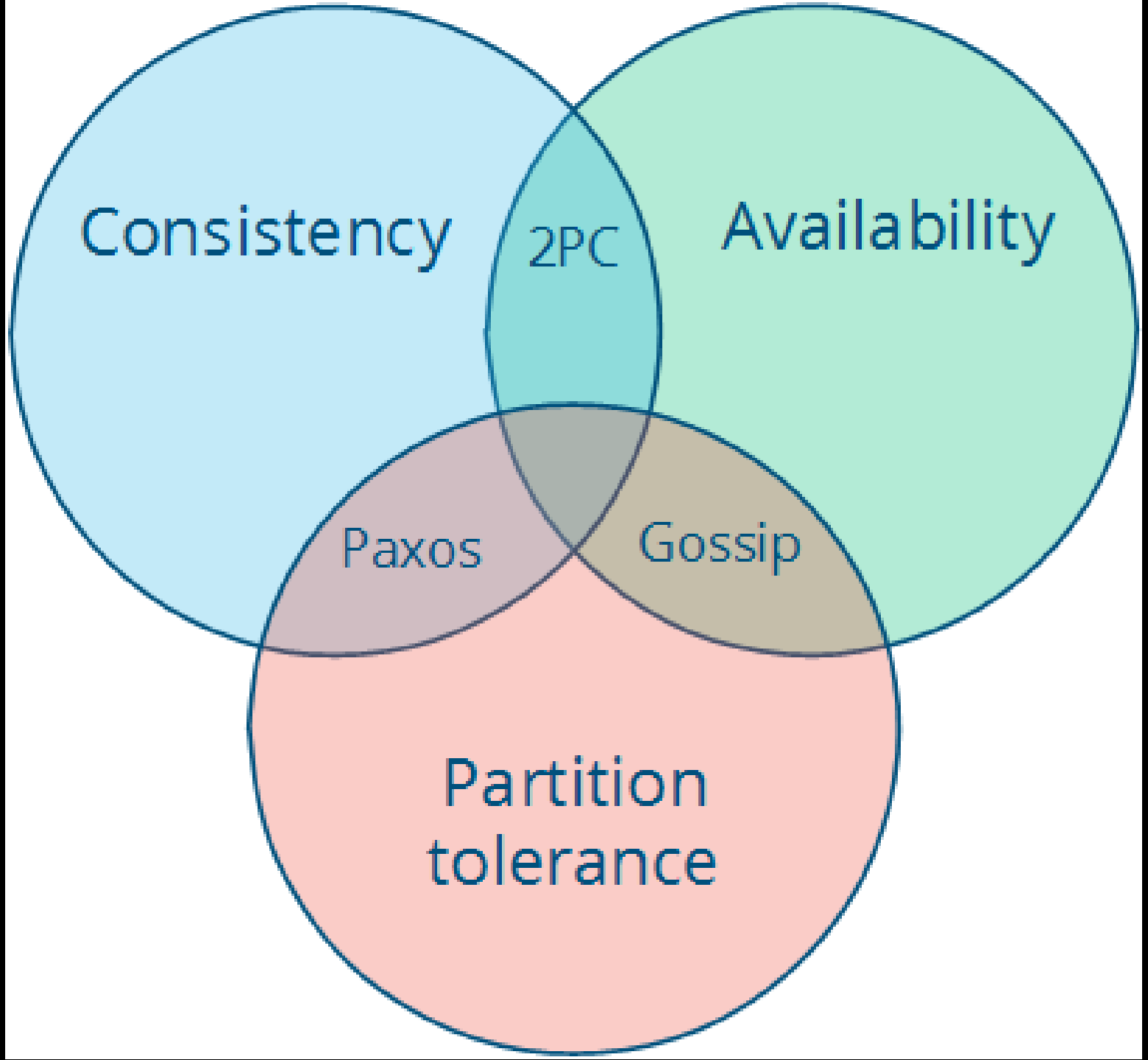
<- ACK

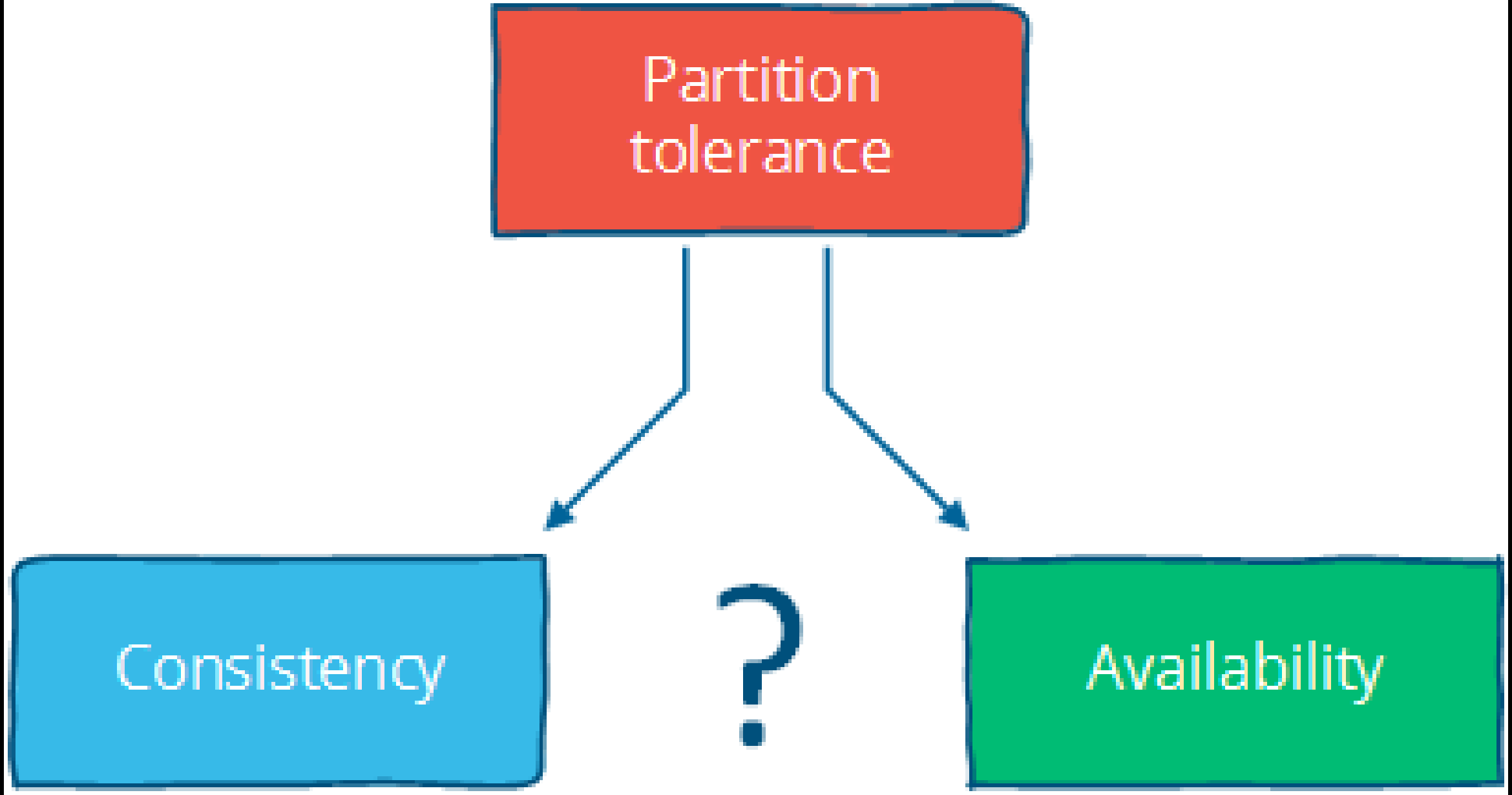
Paxos / Raft

- Hlasování většiny
- Dynamický master
- n/2 – 1 nedostupných nodů je OK

CP systémy

- Raft (etcd)
- Paxos (ZooKeeper)


Consistency
2PC
Availability
Paxos
Gossip
Partition
tolerance

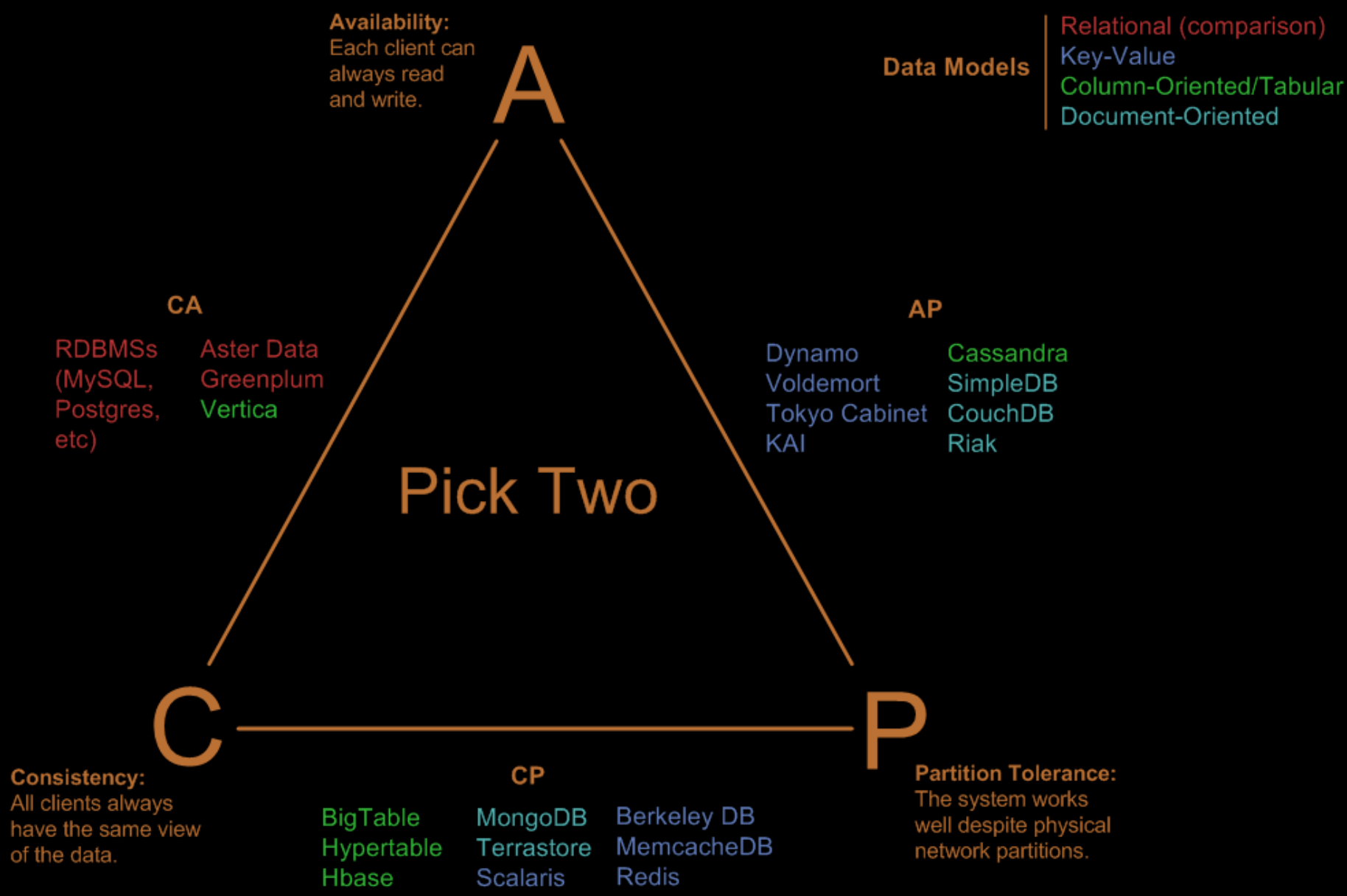

Partition tolerance
Consistency
?
Availability

Děkuju za pozornost.

Štefan Šafár

http://book.mixu.net/distsys/single-page.html

Visual Guide to NoSQL Systems



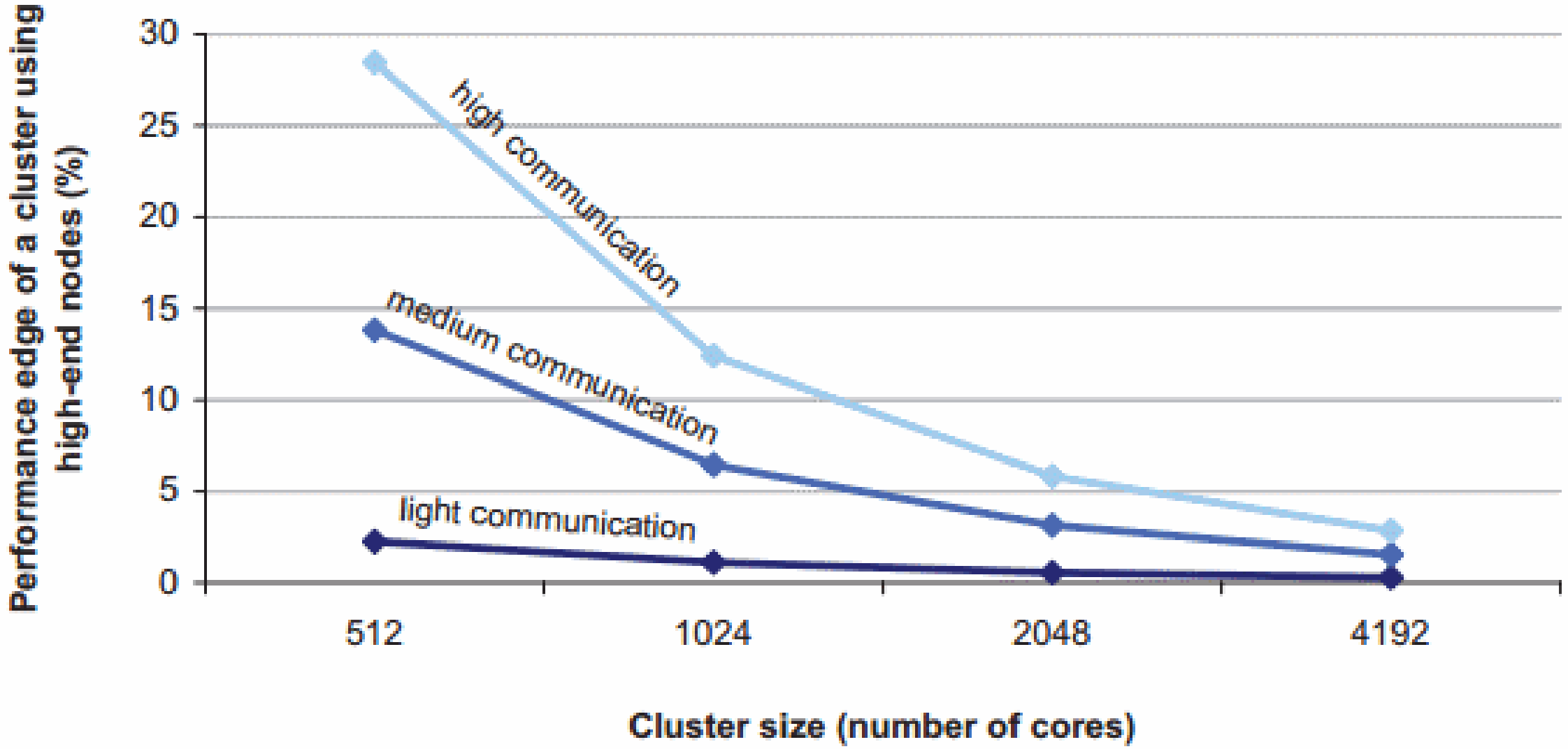


FIGURE 3.2: Performance advantage of a cluster built with high-end server nodes (128-core SMP) over a cluster with the same number of processor cores built with low-end server nodes (four-core SMP), for clusters of varying size.