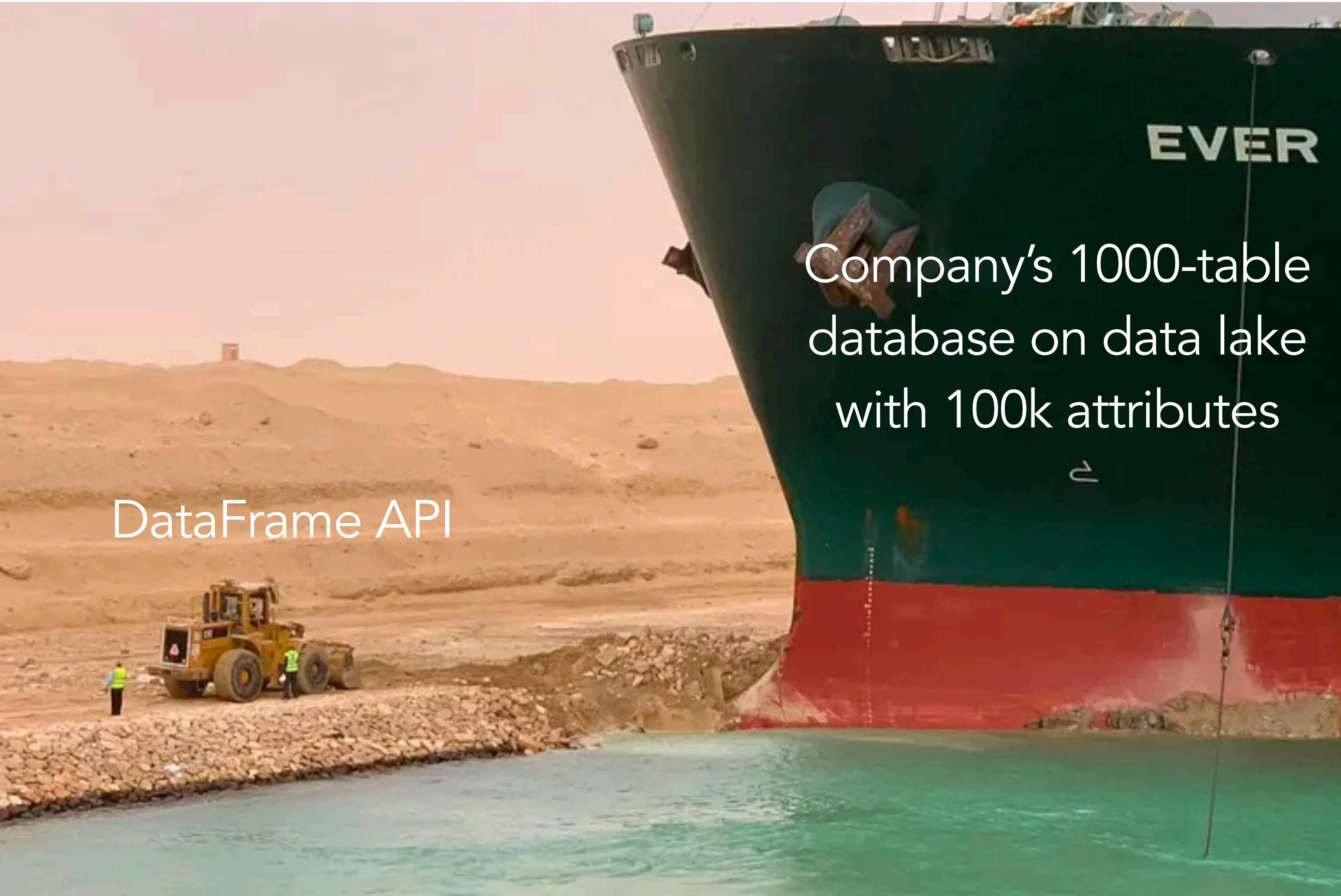


# DSC 204a

## Scalable Data Systems

- Haojian Jin



# Logistics

- PA0 graded.
- PA1 due date.
- Final exam.

# Where are we in the class?

## Foundations of Data Systems (2 weeks)

- Digital representation of Data → Computer Organization → Memory hierarchy → Process → Storage

## Scaling Distributed Systems (3 weeks)

- Cloud → Network → Distributed storage → Parallelism → **Partition and replication**

## Data Processing and Programming model (5 weeks)

- Data Models evolution → Data encoding evolution → → IO & Unix Pipes → Batch processing (MapReduce) → Stream processing (Spark)

# Today's topic: Replication & Partitioning

- PIA
  - Share-X parallelism
  - Replication
  - Partitioning
- 
- Textbook:  
DDIA Ch. 5, 6

# Technologies are always evolving.



## A Case for Redundant Arrays of Inexpensive Disks (RAID)

*David A. Patterson, Garth Gibson, and Randy H. Katz*

Computer Science Division

Department of Electrical Engineering and Computer Sciences

571 Evans Hall

University of California

Berkeley, CA 94720

(pattrsn@ginger berkeley.edu)

*Increasing performance of CPUs and memories will be matched by a similar performance increase in I/O. While Single Large Expensive Disks (SLED) has grown rapidly, the improvement of SLED has been modest. Redundant inexpensive Disks (RAID), based on the magnetic disk developed for personal computers, offers an attractive alternative to SLED, promising improvements of an order of magnitude in reliability, power consumption, and scalability. This paper compares RAID levels to an IBM 3380 and a Fujitsu Super Eagle.*

### 1 Background: Rising CPU and Memory Performance

The users of computers are currently enjoying unprecedented growth in the speed of computers. Gordon Bell said that between 1974 and 1984, single chip computers improved in performance by 40% per year, about twice the rate of minicomputers [Bell 84]. In the following year Bill Joy

ure of magnetic disk technology is the growth in the maximum number of bits that can be stored per square inch, or the bits per inch. In fact, this times the number of tracks per inch. Called M A D , for maximum are density, the "First Law in Disk Density" predicts [Frank87]

$$MAD = 10^{(Year-1971)/10}$$

Magnetic disk technology has doubled capacity and halved price every years, in line with the growth rate of semiconductor memory. In practice between 1967 and 1979 the disk capacity of the average IBM processing system more than kept up with its main memory [Steve

Capacity is not the only memory characteristic that must grow rapidly to maintain system balance, since the speed with which instructions and data are delivered to a CPU also determines its ultimate performance. The speed of main memory has kept pace for two reasons:

- (1) the invention of caches, showing that a small buffer can be automatically to contain a substantial fraction of memory references

# Technologies are always evolving.

## Market Trends—USB Flash Drives, Worldwide, 2001–2010

### Average Selling Price of USB Flash Drives, Worldwide, 2001-2010 (Dollars)—Actual and Projected

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	CAGR 2004-2010
8MB	11.13	8.20	9.75	-	-	-	-	-	-	-	NA
16MB	16.14	11.37	9.51	6.85	6.18	-	-	-	-	-	-100%
32MB	25.68	17.41	12.22	8.04	6.36	5.54	-	-	-	-	-100%
64MB	41.53	27.69	19.52	14.01	8.08	6.37	5.08	4.34	-	-	-100%
128MB	78.18	50.16	29.95	22.16	12.05	8.21	6.01	5.50	4.58	4.07	-25%
256MB	156.83	98.50	51.11	33.10	19.30	12.44	7.69	6.72	5.94	4.92	-27%
512MB	326.34	196.13	112.77	57.79	31.04	20.17	11.58	8.92	7.36	6.69	-30%
1GB	-	410.74	214.31	88.48	55.91	32.69	18.66	14.00	9.95	8.56	-32%
2GB	-	-	3,335.43	166.56	102.86	59.19	30.14	23.28	15.91	11.94	-36%
4GB	-	-	-	-	224.41	109.25	54.44	38.30	26.79	19.72	NA
8GB	-	-	-	-	555.88	238.81	100.34	70.11	44.40	33.92	NA
16GB	-	-	-	-	-	592.17	219.14	130.19	81.70	56.92	NA
32GB	-	-	-	-	-	1,416.67	543.15	285.70	152.16	105.63	NA
64GB	-	-	-	-	-	-	1,212.76	709.82	334.51	197.64	NA
<b>Average</b>	<b>21.09</b>	<b>31.22</b>	<b>29.10</b>	<b>28.40</b>	<b>28.48</b>	<b>25.89</b>	<b>20.94</b>	<b>21.15</b>	<b>19.77</b>	<b>20.71</b>	<b>-5%</b>

Source: Joseph Unsworth, Gartner Dataquest, August 2005 (used with permission).

THE BIG IDEAS BEHIND RELIABLE, SCALABLE,  
AND MAINTAINABLE SYSTEMS

# Designing Data-Intensive Applications

THE BIG IDEAS BEHIND RELIABLE, SCALABLE,  
AND MAINTAINABLE SYSTEMS



Martin Kleppmann

- No silver bullets
  - Lots of trade-off
- Constantly evolving
  - New proposals
    - Succeed → New standard/product
    - Failed → Ideas incorporated
- No black magic
  - Keep revisiting old **BIG** ideas
- Ask me questions!

# Redundancy

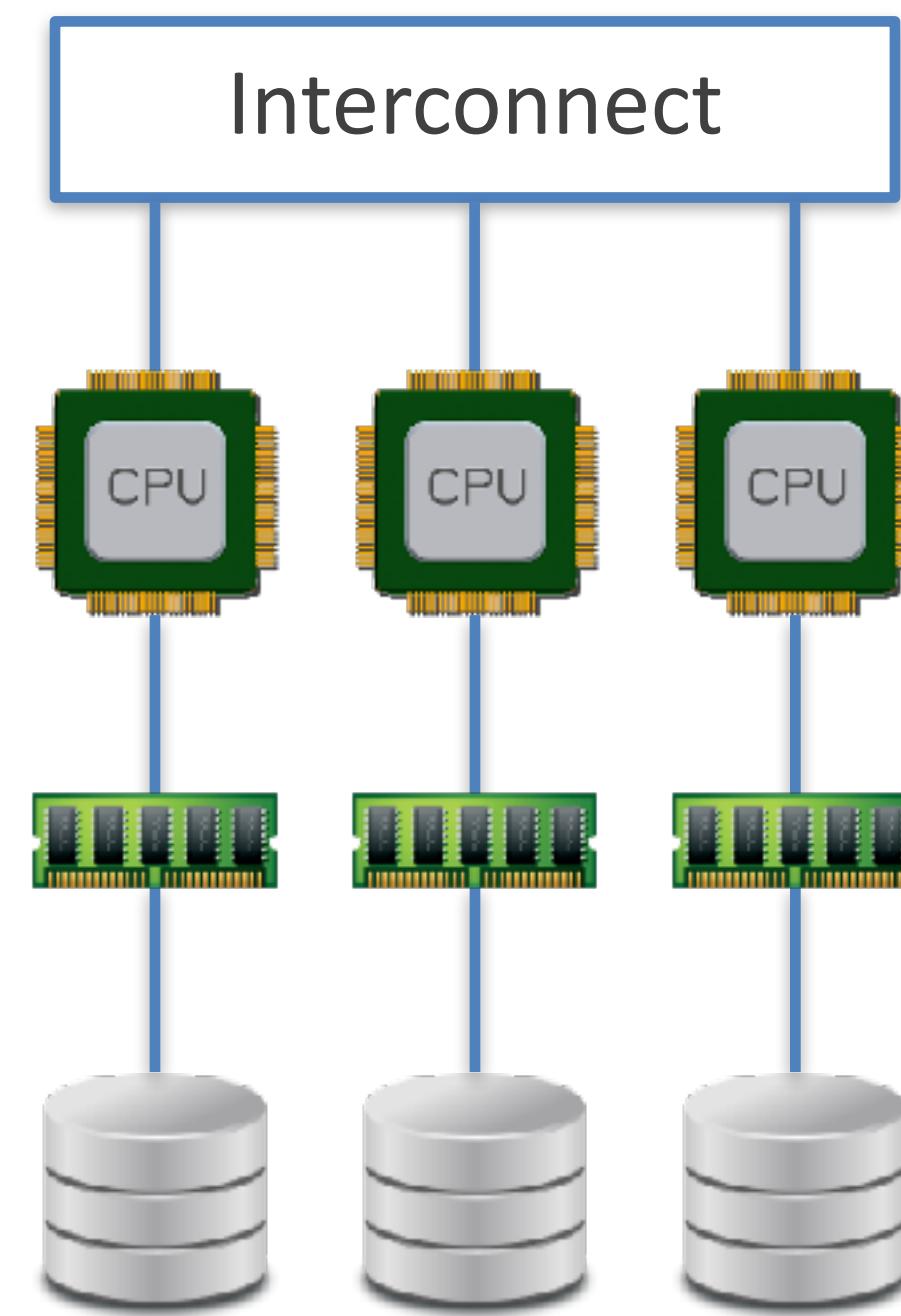
- How many computers are there in the space shuttle?
  - Five general-purpose computers.
    - Four operate during critical mission phase (e.g., ascent, decent)
    - One backup.
  - Four actuators.



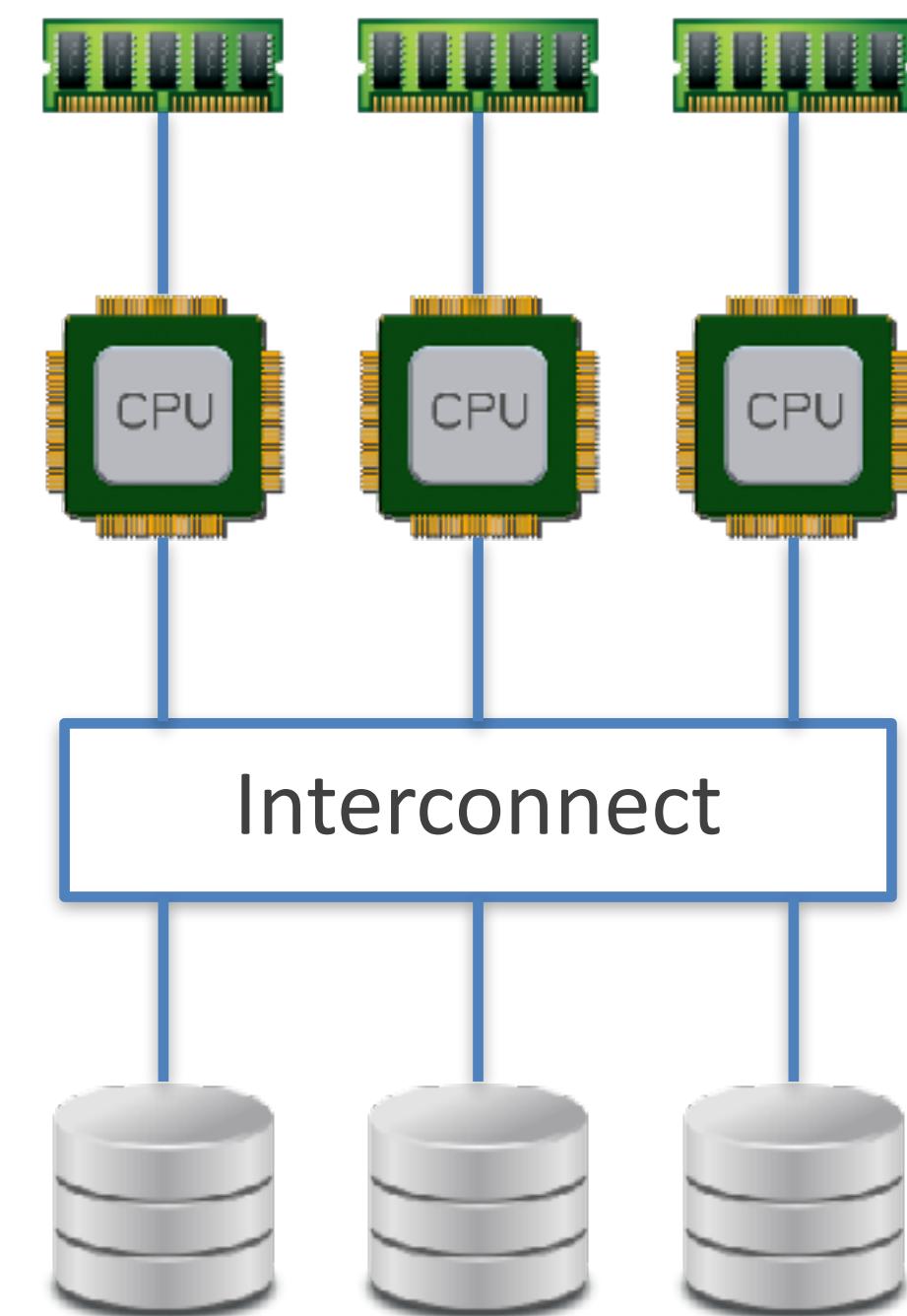
# Why distribute a database across multiple machines?

- Scalability
  - Data volume
  - Read load
  - Write load
- Fault tolerance | high availability
  - When one fails, another can take over.
- Latency
  - Distribute machines worldwide.
  - Reduce network latency.

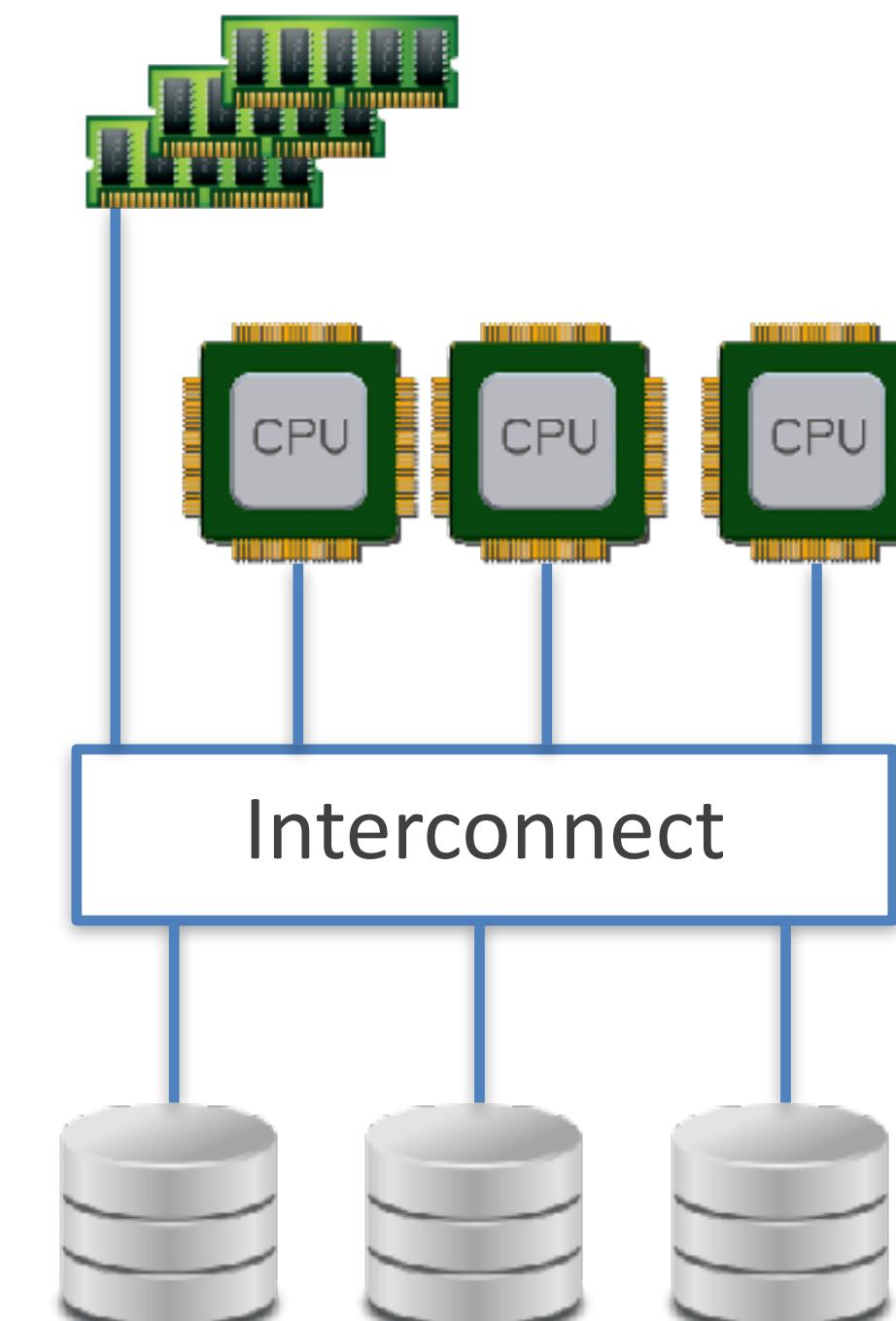
# 3 Paradigms of Multi-Node Parallelism



Shared-Nothing  
Parallelism



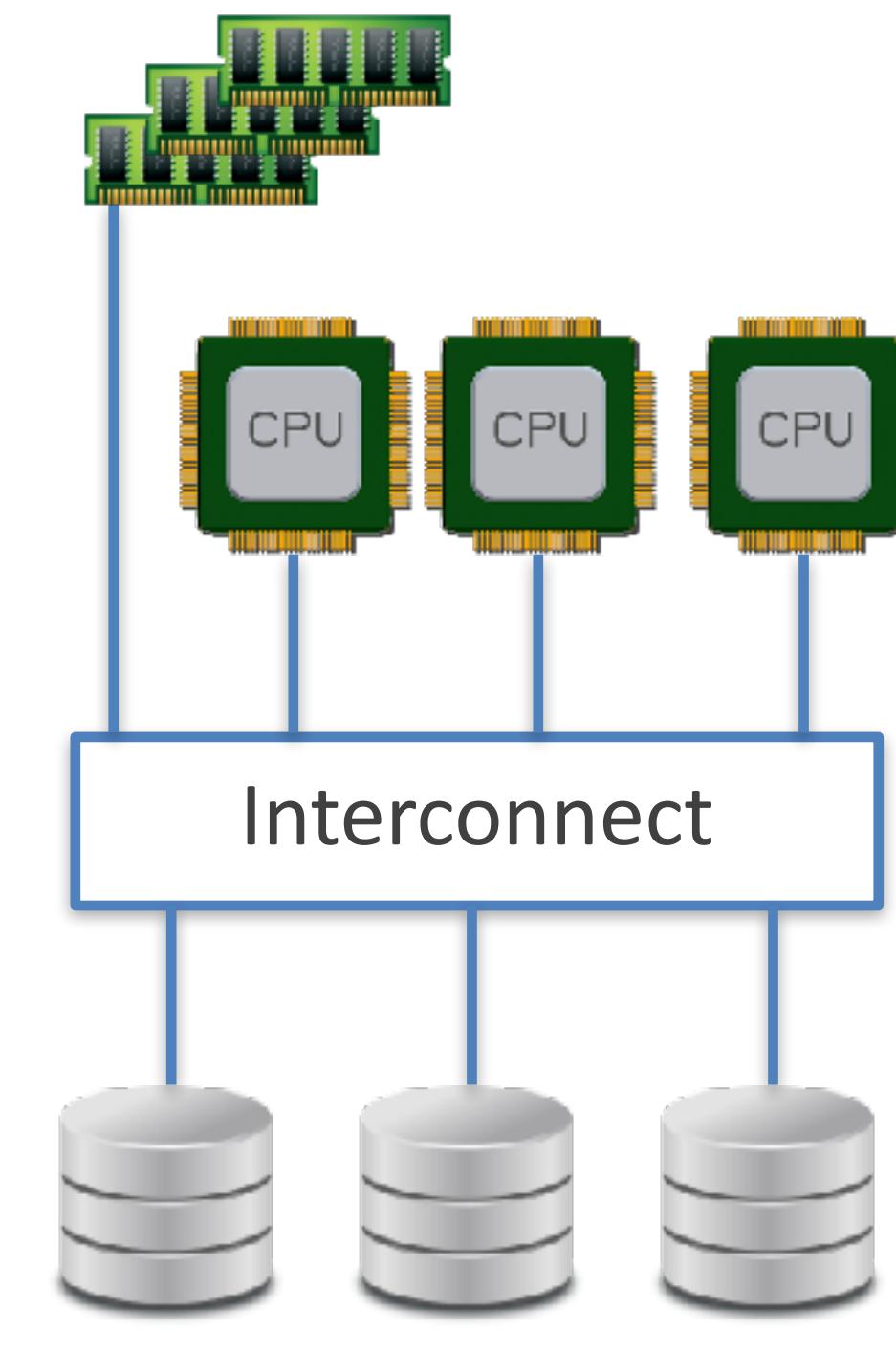
Shared-Disk  
Parallelism



Shared-Memory  
Parallelism

# Shared-memory parallelism (vertical scaling)

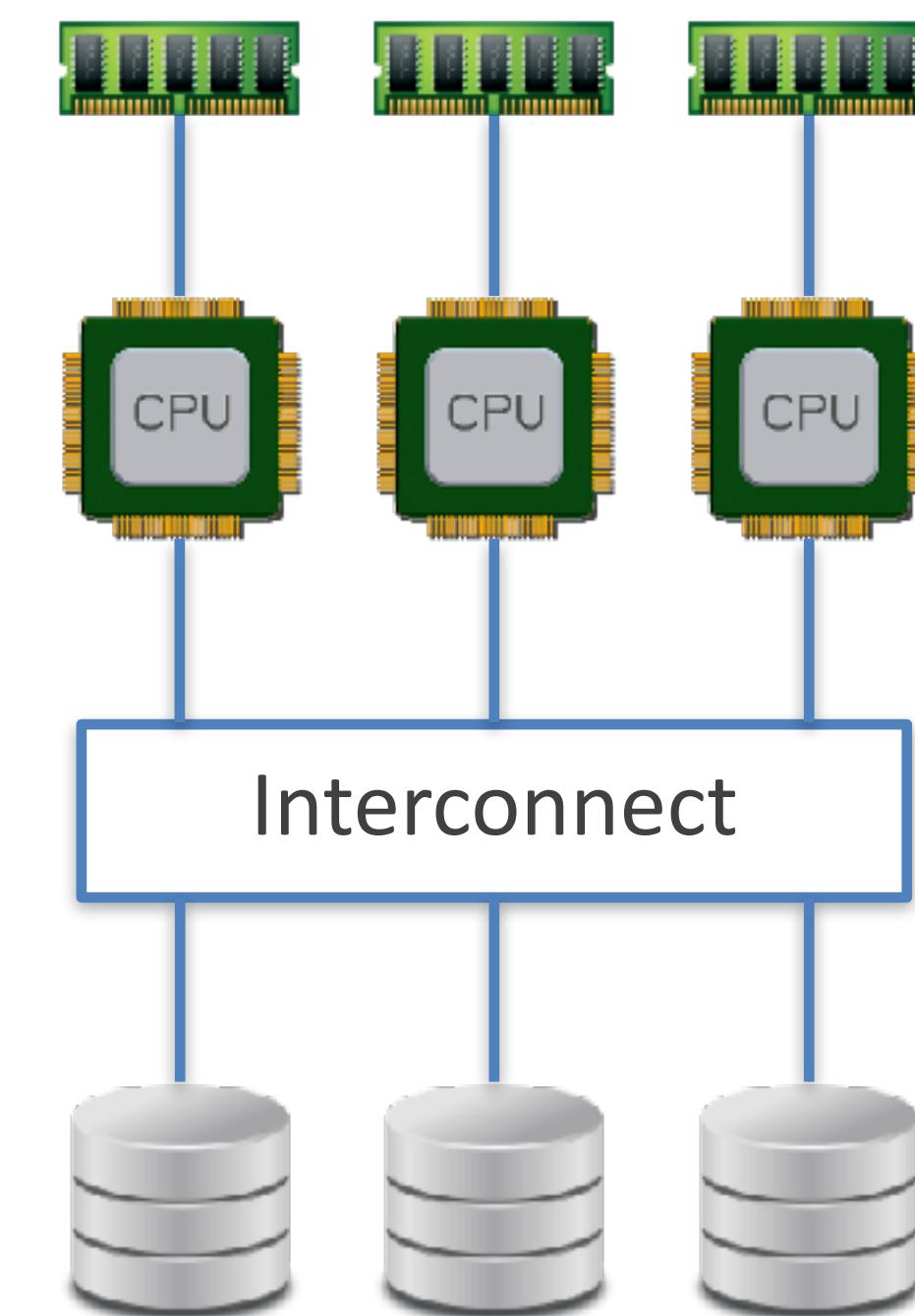
- CPUs, RAM chips, Disks are joined under one OS.
- Advantage:
  - Simple, High performance.
- Disadvantage:
  - Expensive.
    - Cost grows faster super-linearly.
    - Performance grows sub-linearly.
  - Limited fault tolerance
    - Note hot-swappable
    - Geolocation
  - App: Large language model training.



Shared-Memory  
Parallelism

# Shared-Disk Parallelism (data warehouse)

- Machines with independent CPUs and RAM
  - But stores data on an array of disks
- Advantage:
  - Low-cost
- Disadvantage:
  - Overhead of locking limit the scalability



## Synology 2 bay NAS DiskStation DS220+ (Diskless),Black



Roll over image to zoom in



Brand: Synology

4.7 ★★★★★ 3,538 ratings | 322 answered questions

#1 Best Seller in Network Attached Storage (NAS) Enclosures

\$299<sup>99</sup>

FREE Returns

Pay \$25.00/month for 12 months, interest-free upon approval for Amazon Visa

Available at a lower price from other sellers that may not offer free Prime shipping.

Extra Savings 90 days FREE music unli... 1 Applicable Promotion

Size: 2-bay; 2GB DDR4

2-bay; 2GB DDR4

4-bay; 2GB DDR4

4-bay; 4GB DDR4

Style: DS220+

DS220+

DS723+

DS723+ + D4ES01-16G

DS723+ + E10G22-T1-Mini

DS723+ + SNV3410-400G

DS723+ + SSD SNV3410 800GB

DS423+

DS923+

Brand Synology

Color Black

Item Dimensions 6.5 x 4.25 x 9.14 inches

LxWxH

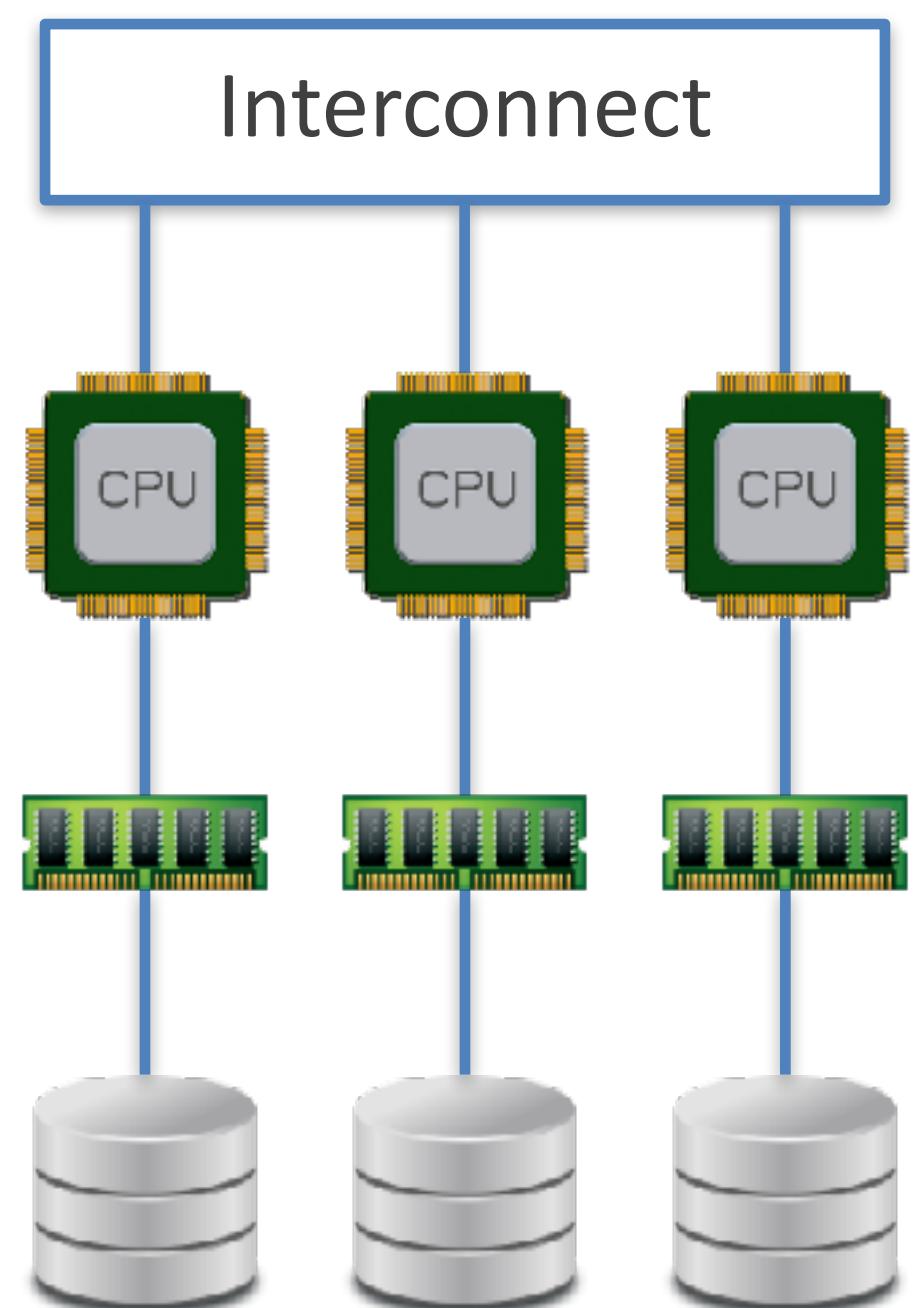
Size 2-bay; 2GB DDR4

Compatible Smartphone, Television, Laptop, Desktop, Camera Devices

### About this item

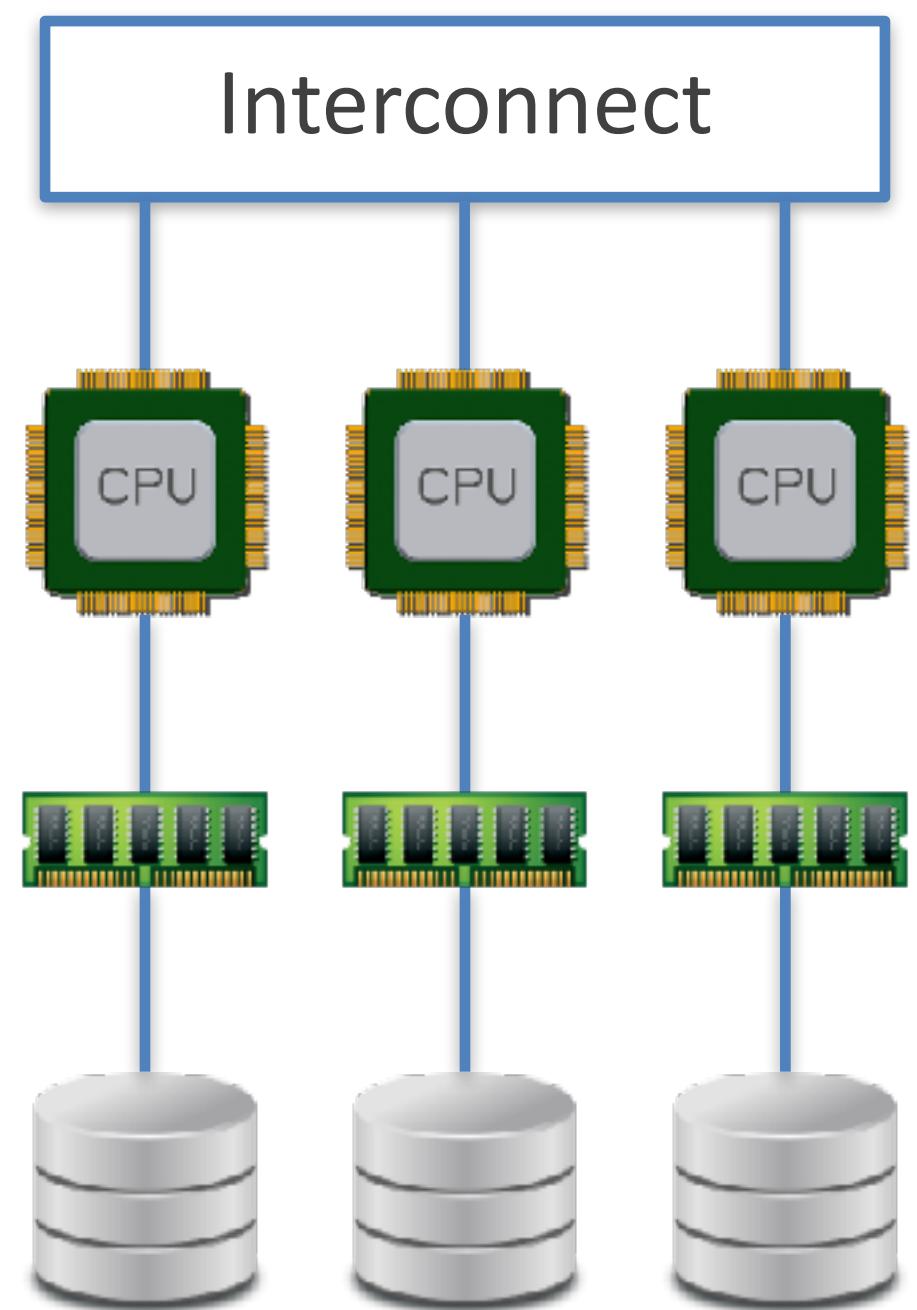
- Featured dual 1GbE LAN ports to support network failover, and with Link Aggregation enabled, DS220+ provides over 225 MB/s sequential read and 192 MB/s sequential write throughput. Data can be further protected with RAID 1 disk mirroring to prevent sudden drive failure.
- Intel dual-core processor with AES-NI hardware encryption engine; 2 GB DDR4 memory (expandable up to 6 GB)

# Shared-Nothing Parallelism (horizontal scaling)



- Most popular approach!
- Each node uses its CPUs, RAM, and disks independently.
  - Any coordination, software level, through a conventional network.
- No silver bullet.
  - Not always the best one.

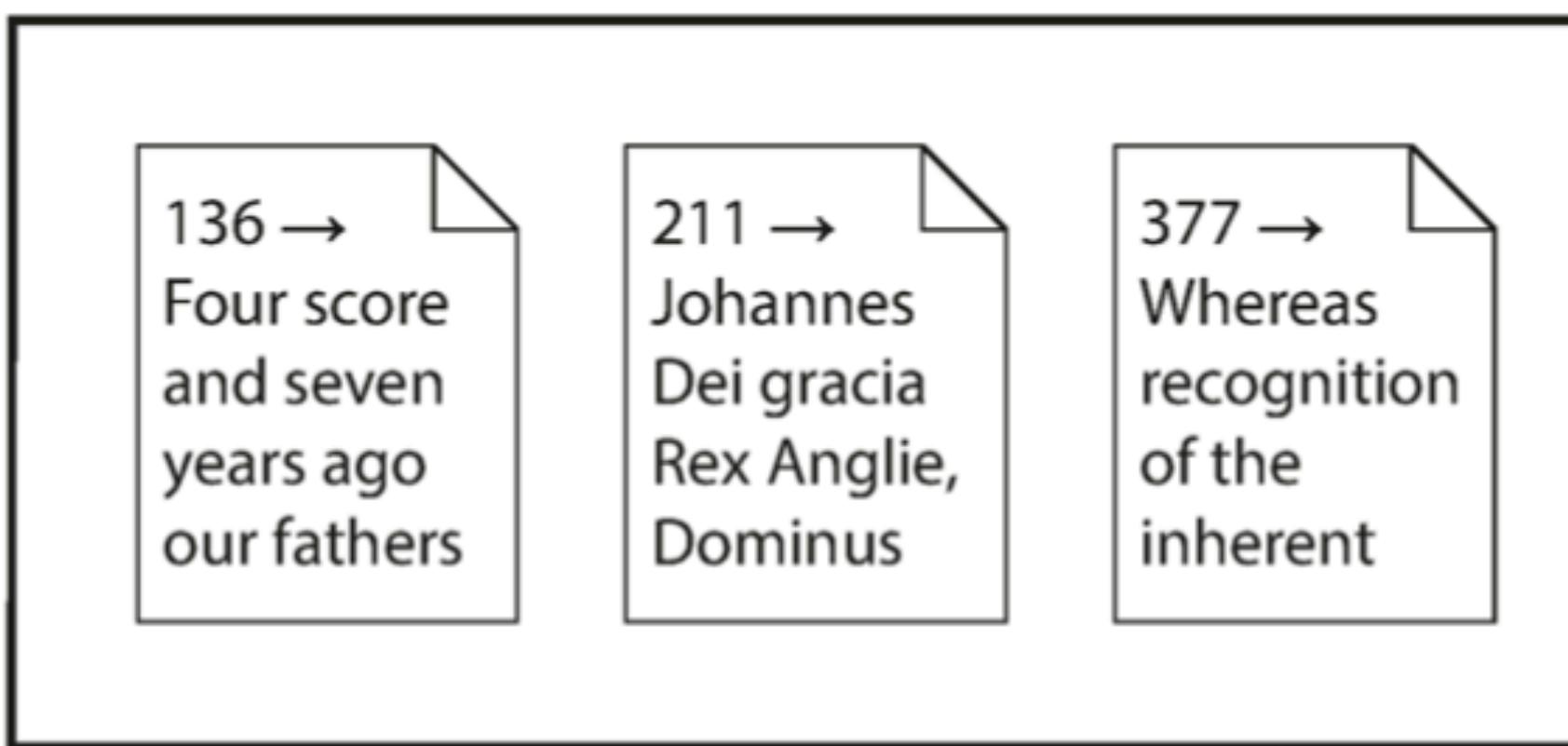
# Shared-Nothing Parallelism (horizontal scaling)



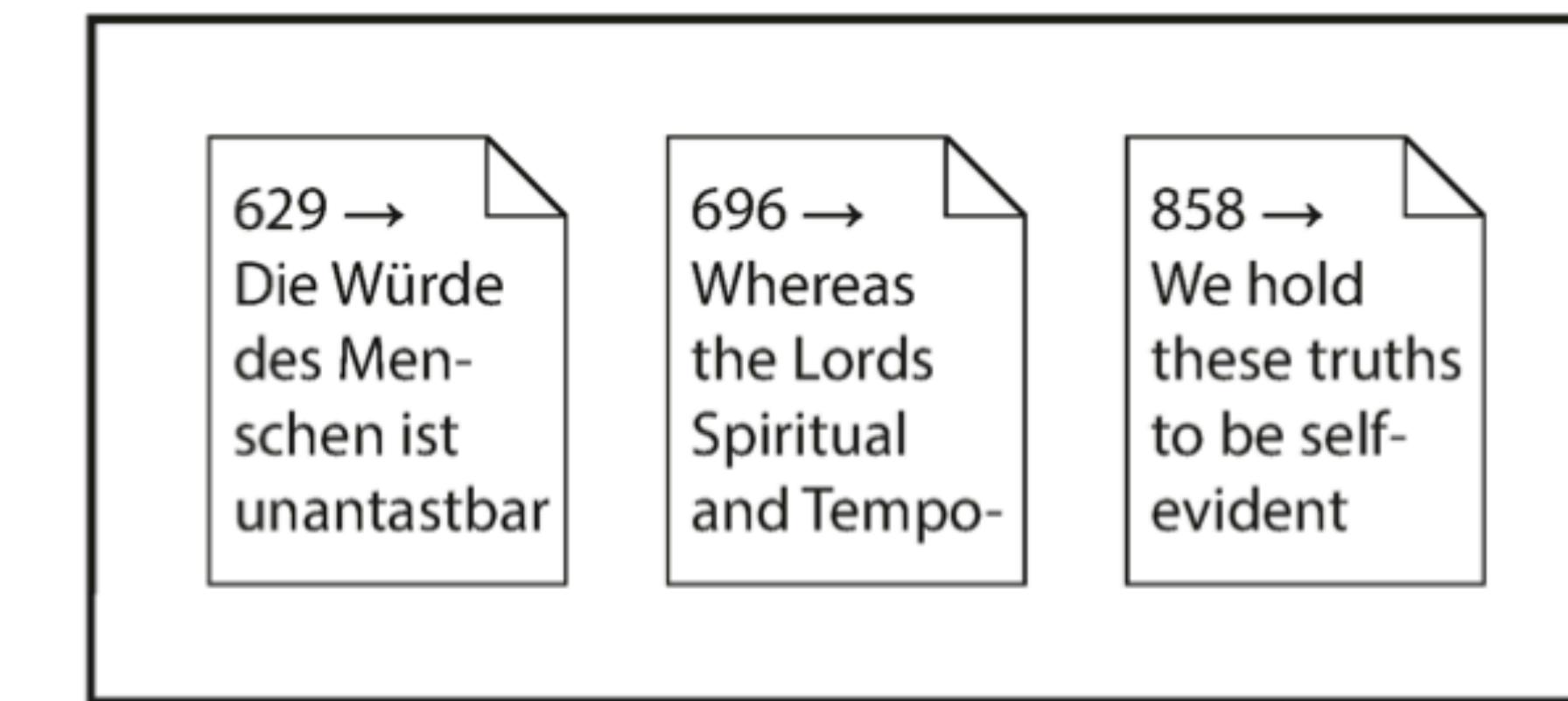
- Advantage:
  - Performance
  - Cost
- Disadvantage
  - Complexity.
  - Involves many constraints and trade-offs.
- Database cannot magically hide all these from you.

# Replication versus Partitioning

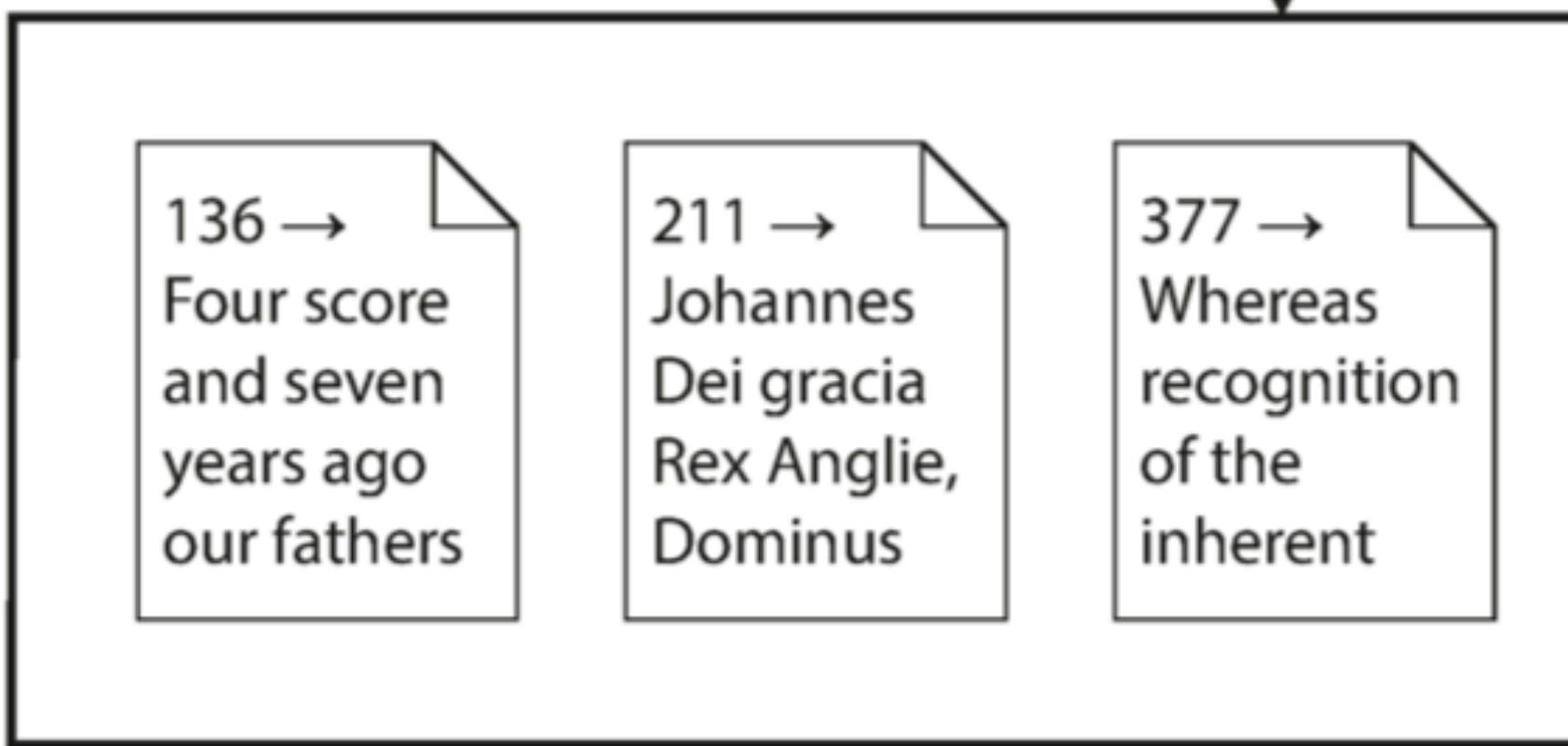
Partition 1, Replica 1



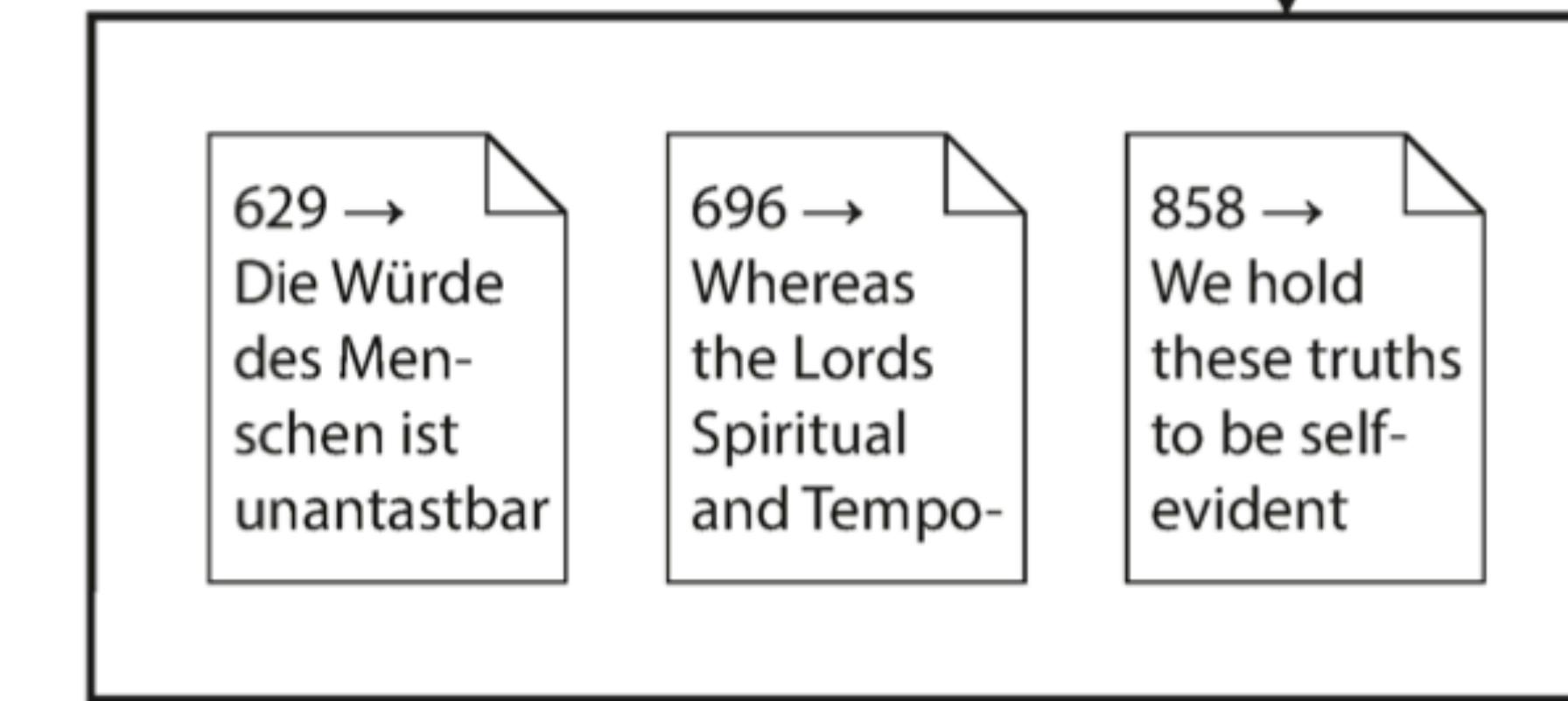
Partition 2, Replica 1



Partition 1, Replica 2



Partition 2, Replica 2



# Today's topic: Replication & Partitioning

- PIA
  - Share-X parallelism
  - **Replication**
  - Partitioning
- 
- Textbook:  
DDIA Ch. 5, 6

Recall:

## Why distribute a database across multiple machines?

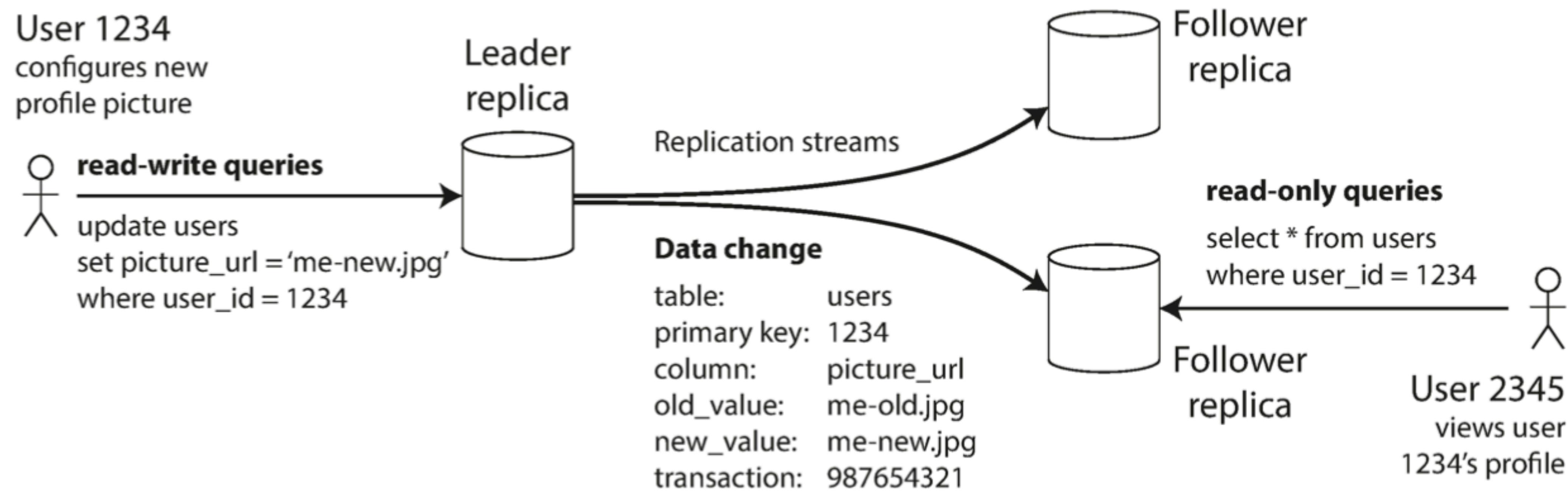
- Scalability
  - Data volume
  - Read load
  - Write load
- Fault tolerance | high availability
  - When one fails, another can take over.
- Latency
  - Distribute machines worldwide.
  - Reduce network latency.

# Replication (assuming a small dataset)

- Scalability
  - Data volume
  - Read load
  - Write load
- Fault tolerance | high availability
  - When one fails, another can take over.
- Latency
  - Distribute machines worldwide.
  - Reduce network latency.

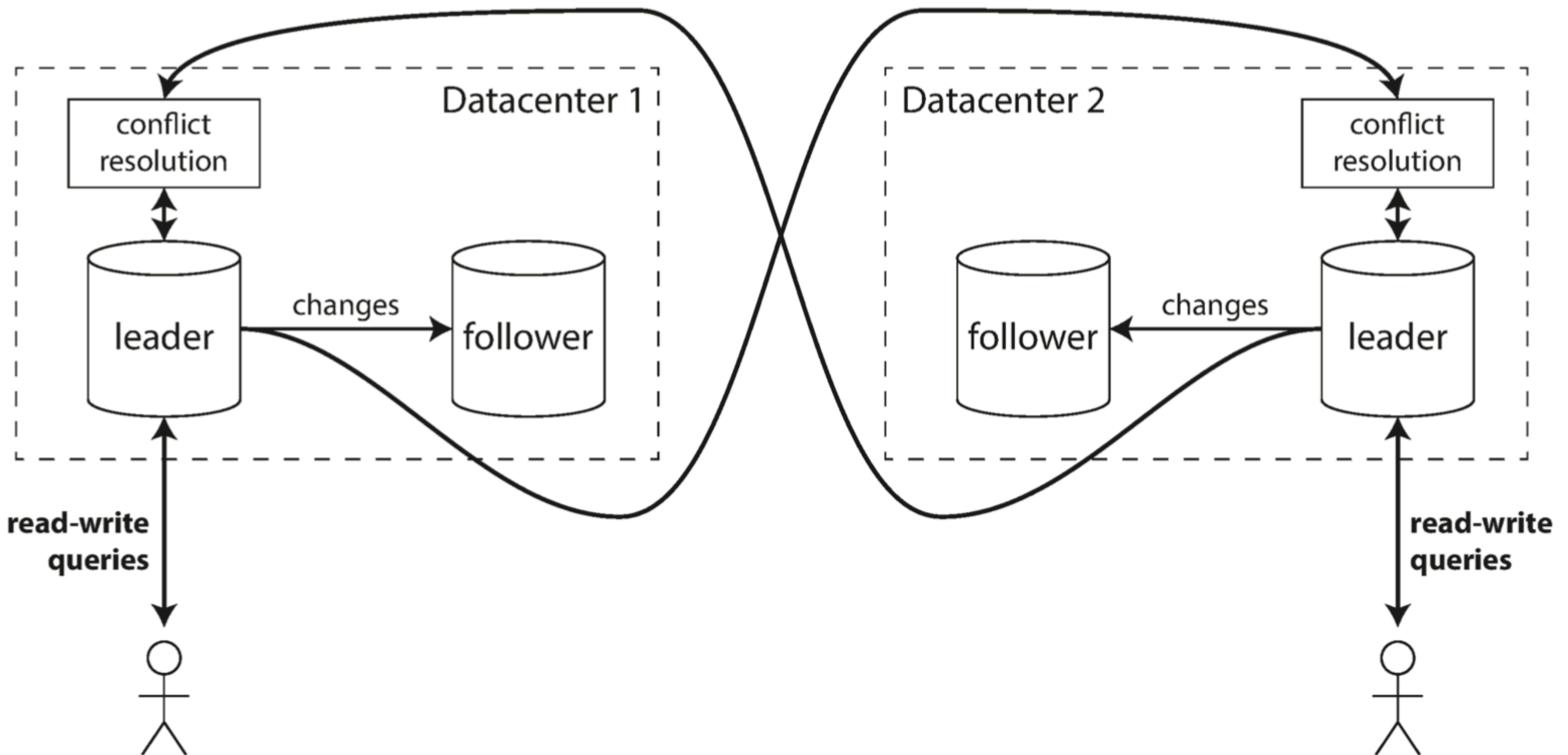
# Challenge: How to handle changes to replicated data?

# Leaders and Followers



- Clients send requests to the leader to write to the db. The leader saves locally
- The leader then sends the data change to all of its followers.
- Clients read data from the leader or followers.
- One of the big ideas in CS (master-slave). Distributed msg brokers (e.g., Kafka).

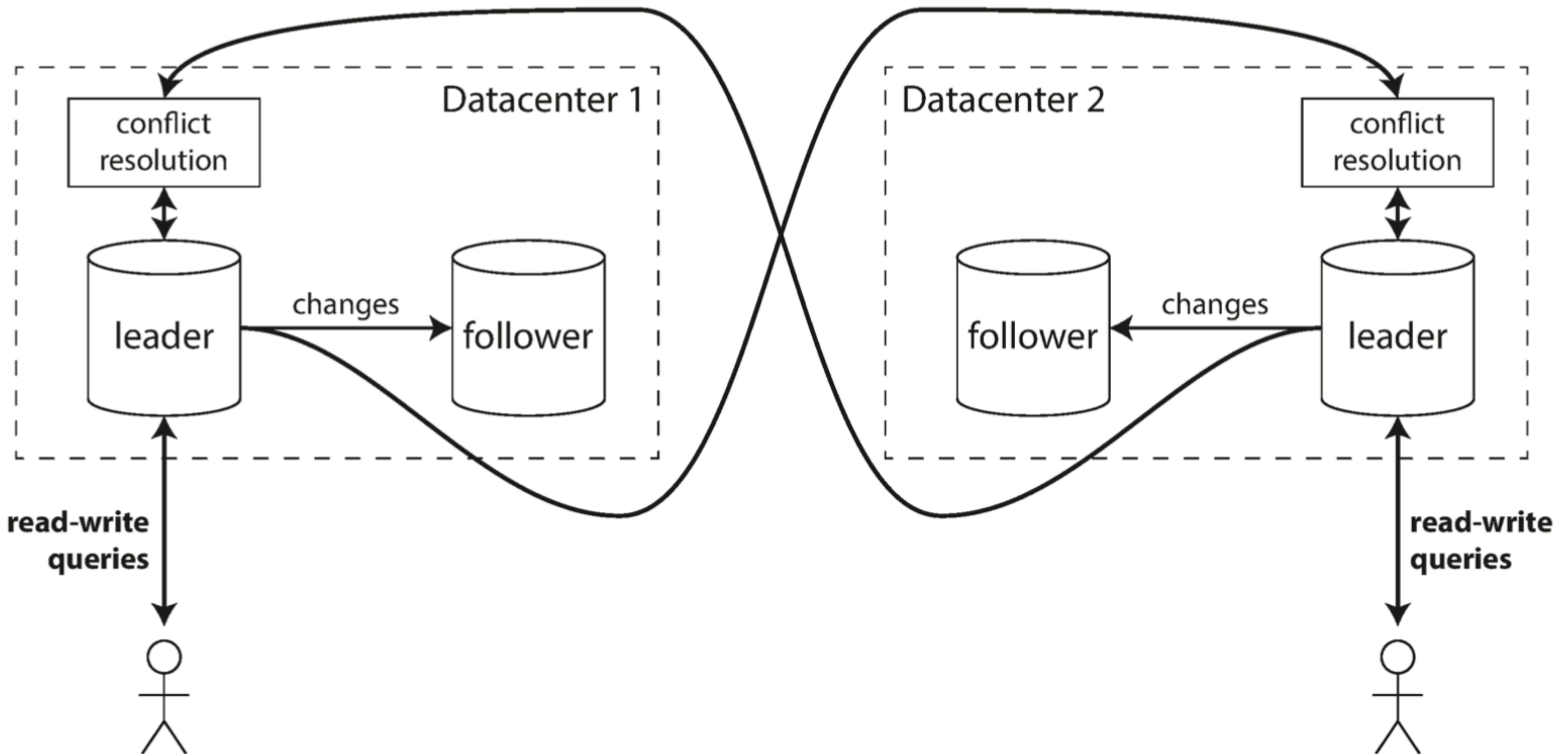
# Single-leader replication



# Single-leader replication

- Advantages:
  - Simplicity (easy-to-understand)
  - Conflicts across nodes (consistency)
- Disadvantages:
  - Single point of failure

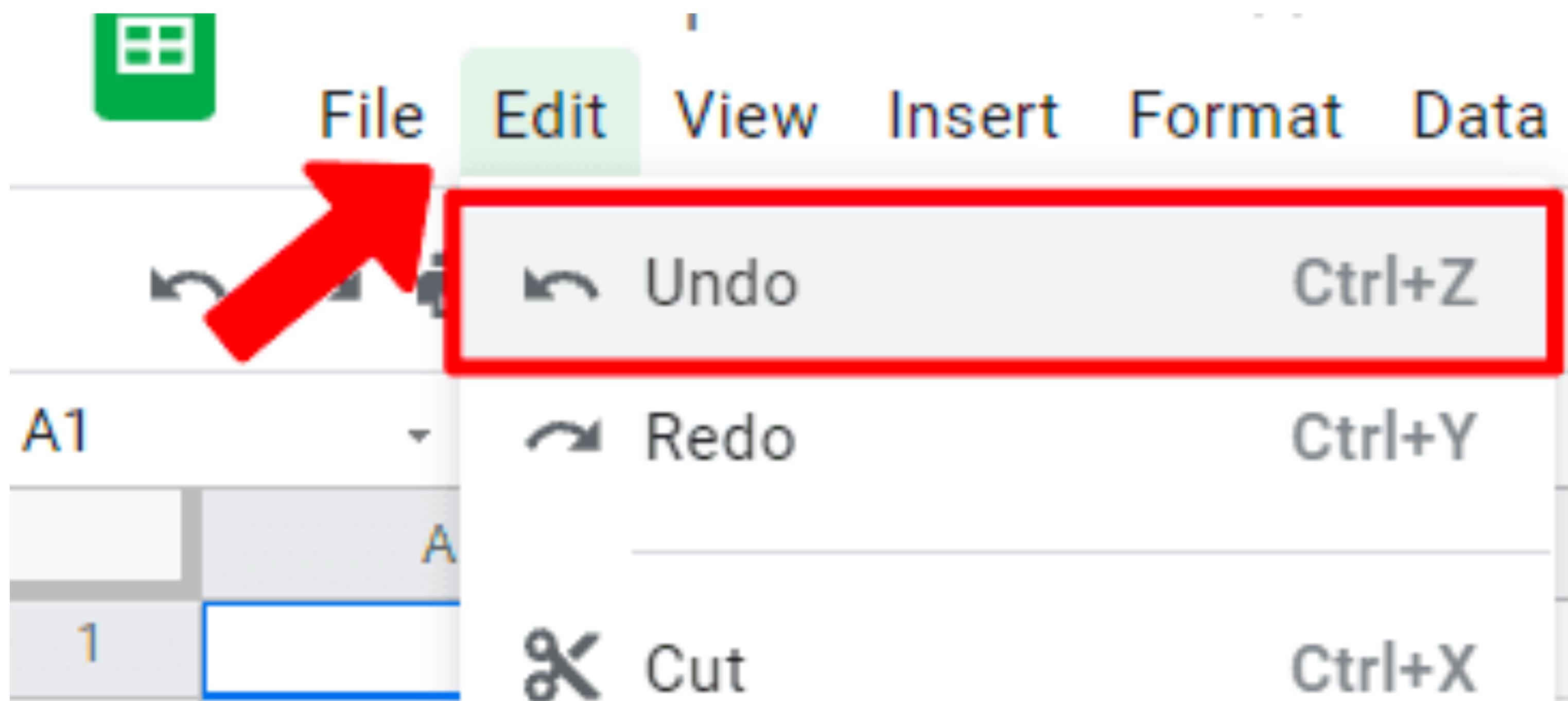
# Multi-leader replication (Multi-datacenter operation)



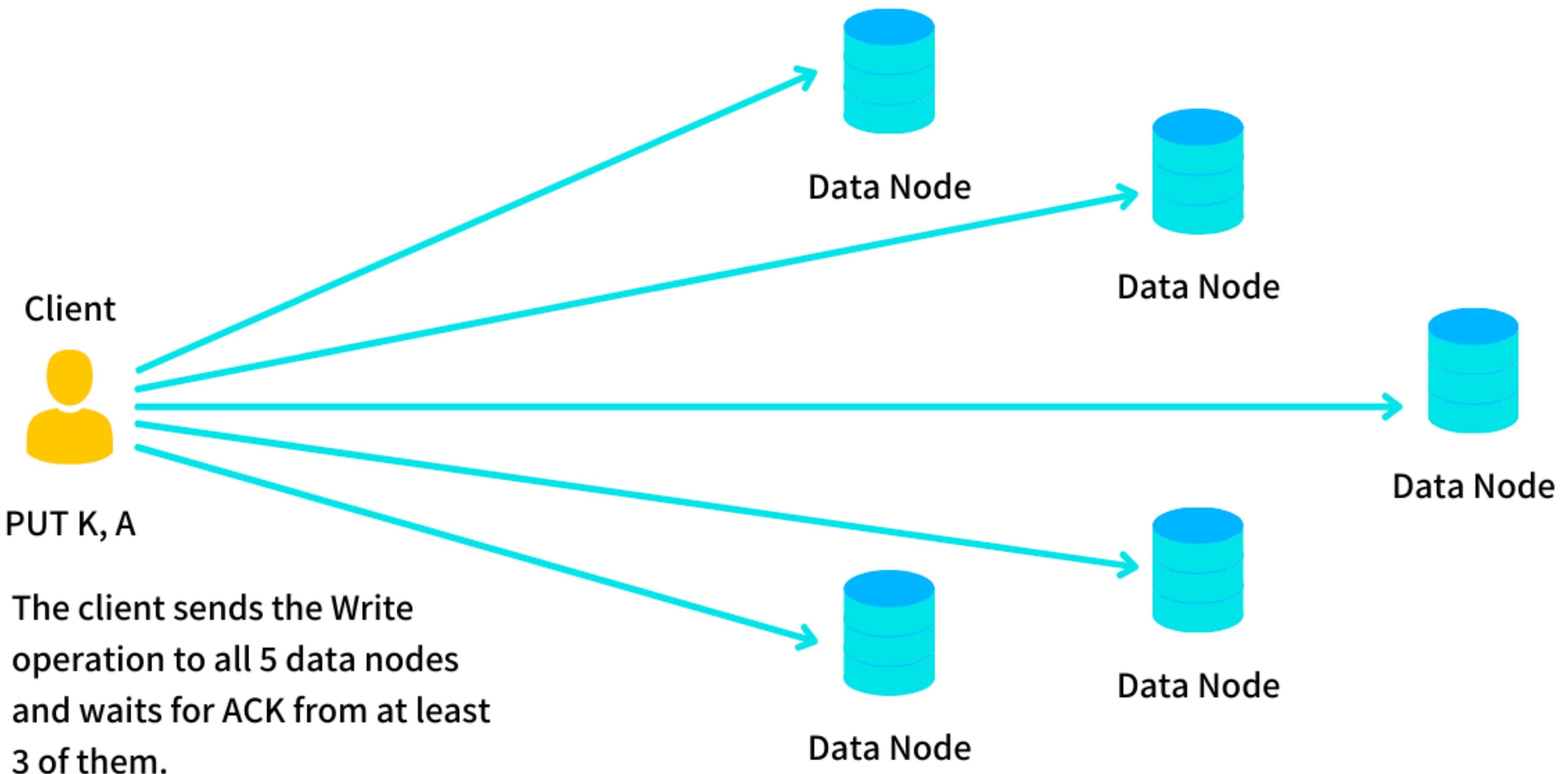
# Multi-leader replication

- Advantages:
  - Performance (i.e., network latency).
  - Tolerance of data center outages.
  - Tolerance of network problems.
    - Asynchronous syncing.
- Disadvantages:
  - Potential write conflicts.
    - The same data may be concurrently modified in two different data centers.
  - Only use it if necessary.

# Google Doc Undo-Redo



# Leaderless replication



# Challenge: How to handle changes to replicated data?

- Three main approaches:
  - Single-leader replication
  - Multi-leader replication
  - Leaderless replication
- Tradeoffs
  - Simplicity (easy-to-understand)
  - Conflicts across nodes (consistency)
  - Faulty nodes
  - Network interruption
  - Latency spikes

# Replication (assuming a small dataset)

- Scalability
  - Data volume
  - Read load
  - Write load
- Fault tolerance | high availability
  - When one fails, another can take over.
- Latency
  - Distribute machines worldwide.
  - Reduce network latency.

# Partitioning (what if the dataset is too big for a single machine?)

## **Bigtable: A distributed storage system for structured data**

[PDF] acm.org

F Chang, J Dean, S Ghemawat, WC Hsieh... - ACM Transactions on ..., 2008 - dl.acm.org

... Despite these varied demands, **Bigtable** has ... **Bigtable**, which gives clients dynamic control over data layout and format, and we describe the design and implementation of **Bigtable**...

☆ Save 99 Cite Cited by 7813 Related articles All 229 versions

## [PDF] The **Google file system**

[PDF] acm.org

S Ghemawat, H Gobioff, ST Leung - ... symposium on Operating systems ..., 2003 - dl.acm.org

... the **Google File System**, a scalable distributed **file system** for ... goals as previous distributed **file systems**, our design has ... departure from some earlier **file system** assumptions. This has led ...

☆ Save 99 Cite Cited by 10064 Related articles All 305 versions

## **MapReduce: simplified data processing on large clusters**

[PDF] acm.org

J Dean, S Ghemawat - Communications of the ACM, 2008 - dl.acm.org

... pairs, and then applying a **reduce** operation to all the values ... with user-specified **map** and **reduce** operations allows us to ... implementation of the **Map Reduce** interface tailored towards ...

☆ Save 99 Cite Cited by 22507 Related articles All 96 versions

# Combining replication and partitioning

