The Secret Life of Consensus in Distributed Systems

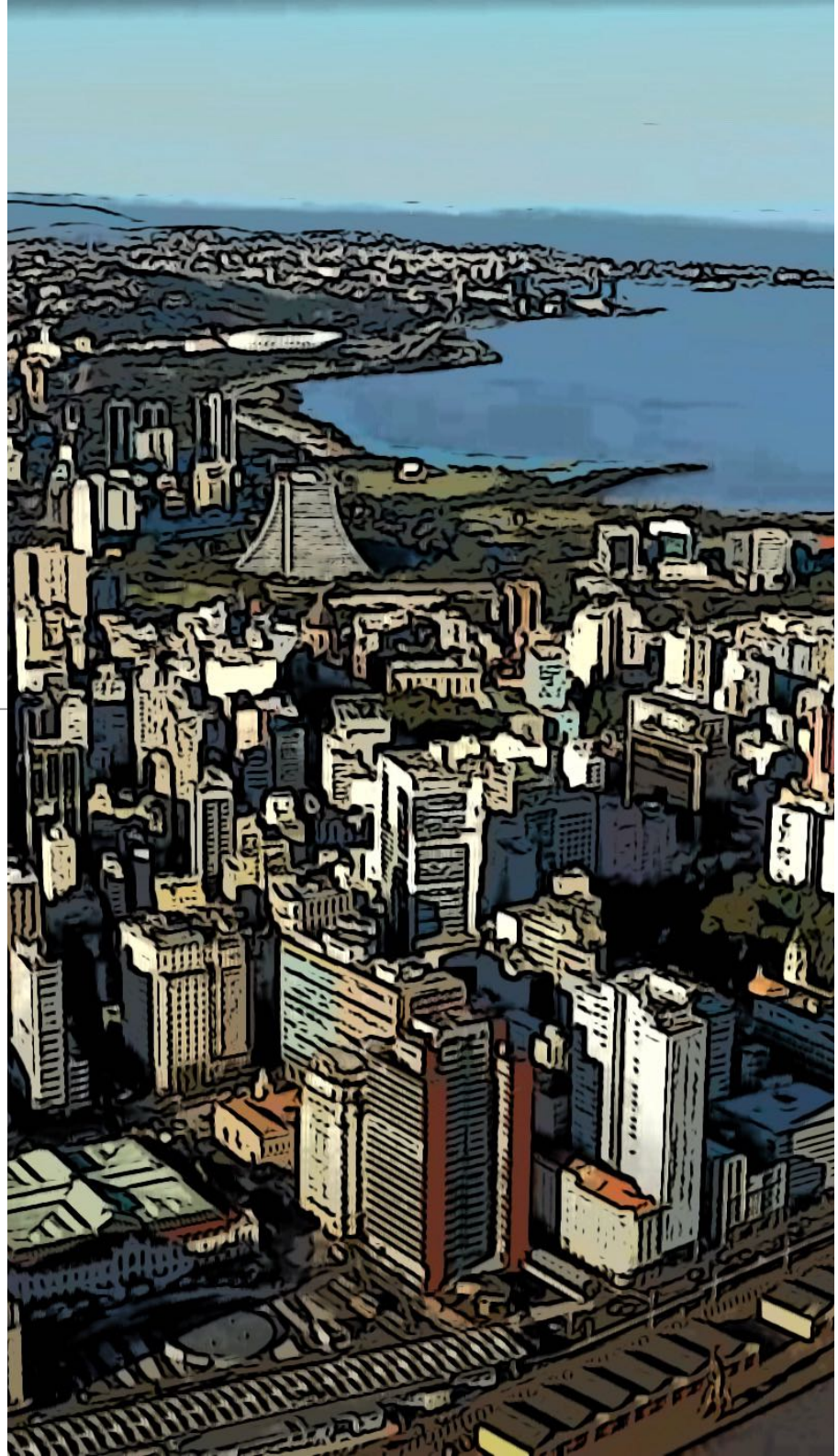
—What Blockchains, Telecoms, and Banks Have in Common

Fernando Pedone

Università della Svizzera italiana (USI)
Switzerland

Workshop Suíça–Brasil @ PUCRS

Porto Alegre, April 15-16, 2025

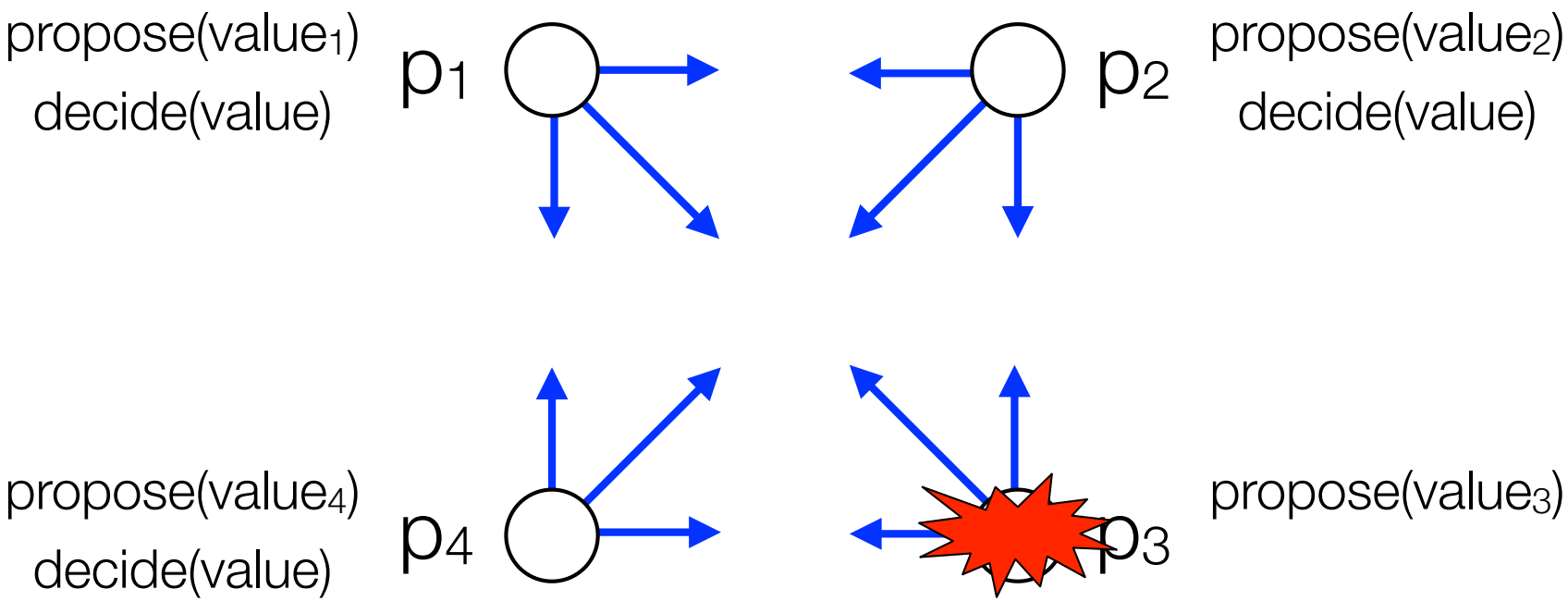


Roadmap

- The notion and history of consensus
- Case study 1: Large-scale graph processing
- Case study 2: Database replication
- Case study 3: Blockchain
- The future of consensus

The consensus problem

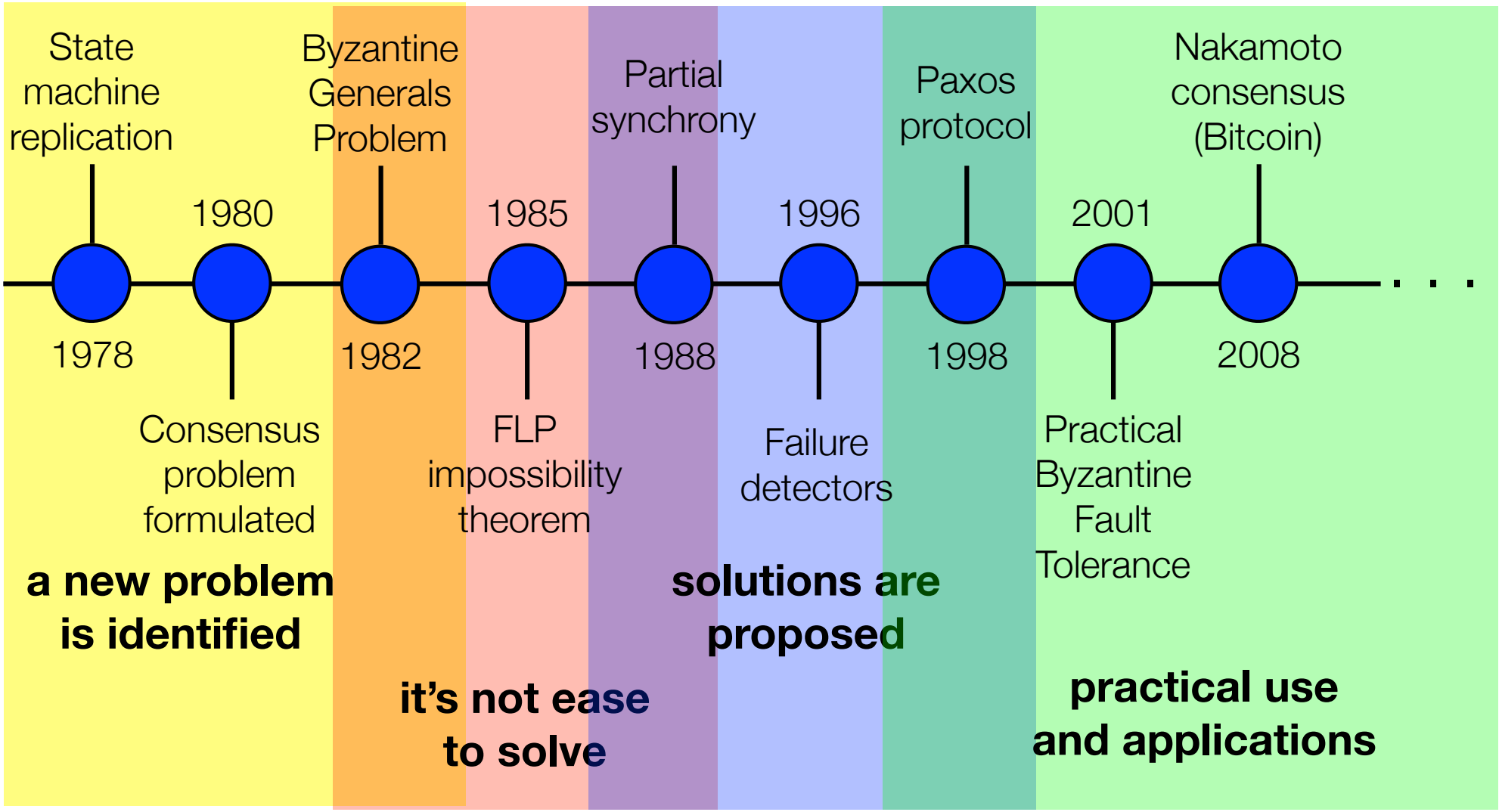
- Processes **propose** values and **decide** on a single value
- Difficult due to failures, message loss, limited synchrony



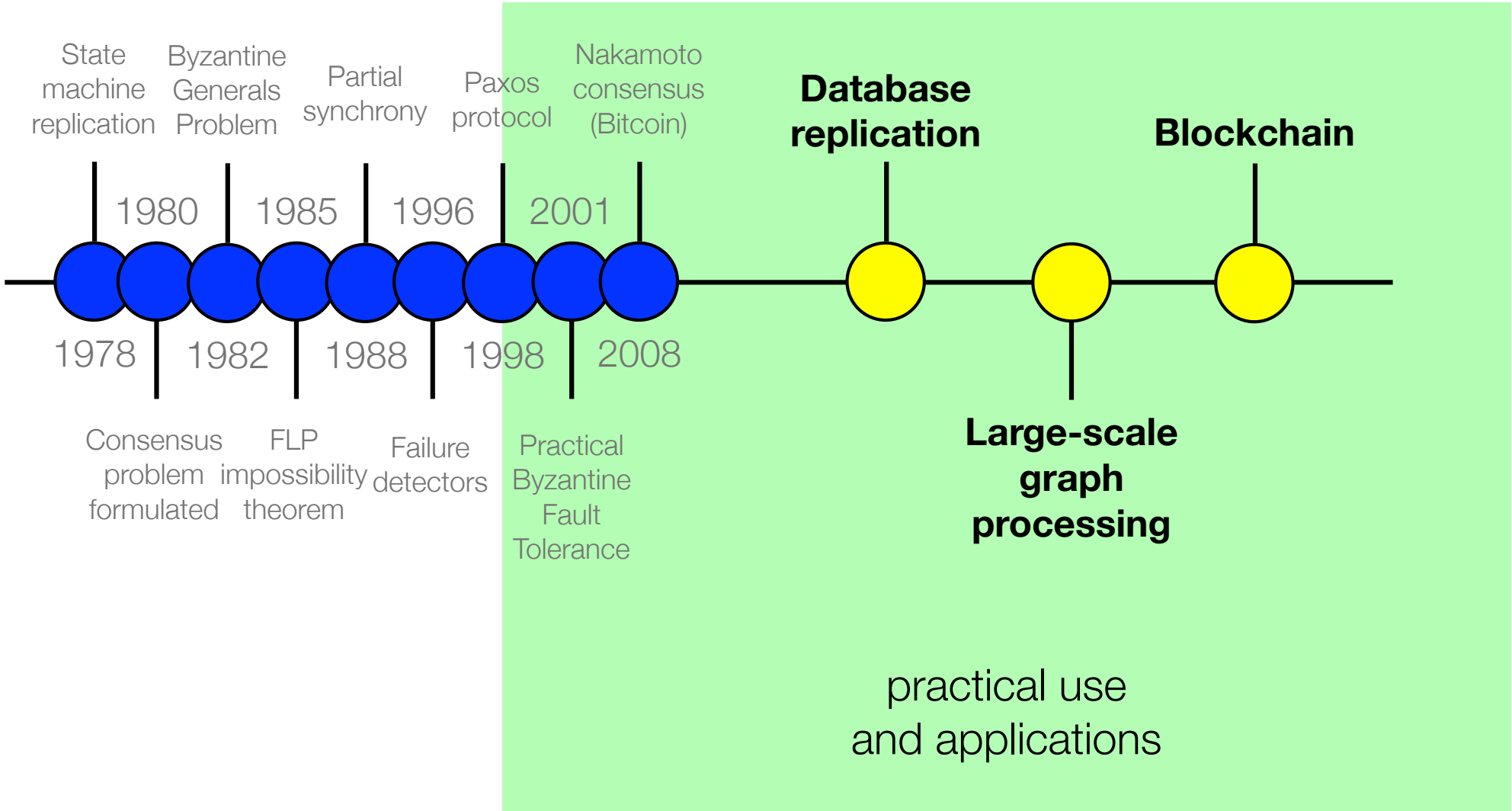
The consensus problem

- Defined by three properties
  - ✦ Processes decide on the same value v
  - ✦ If a process decides v, then v was proposed
  - ✦ Every non-faulty process eventually decides

Consensus timeline

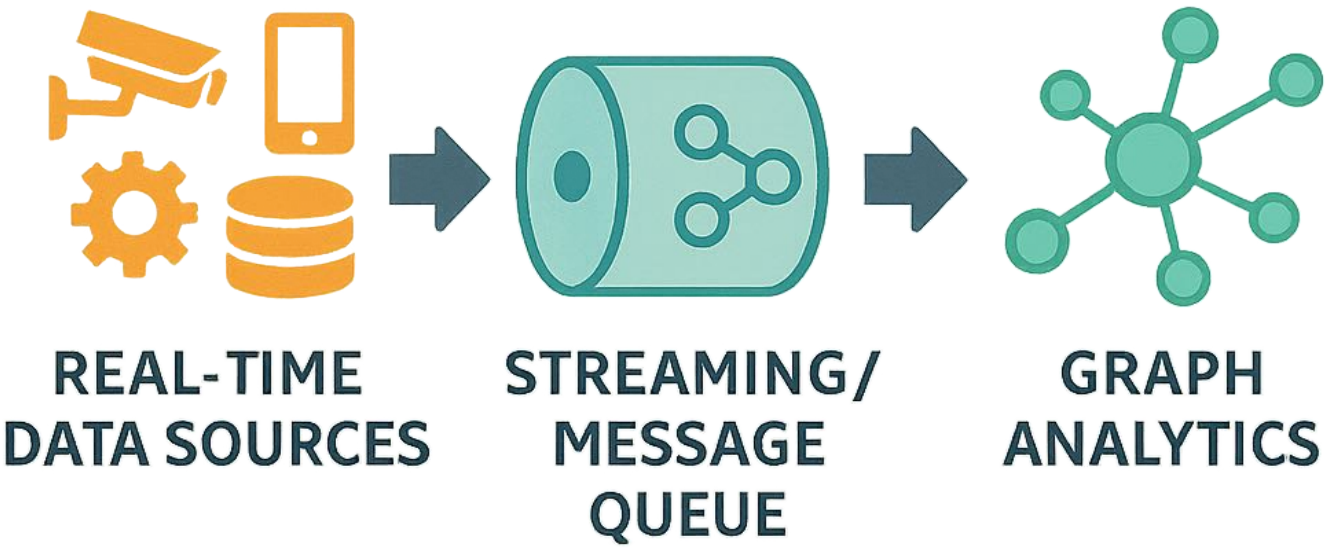


Consensus today



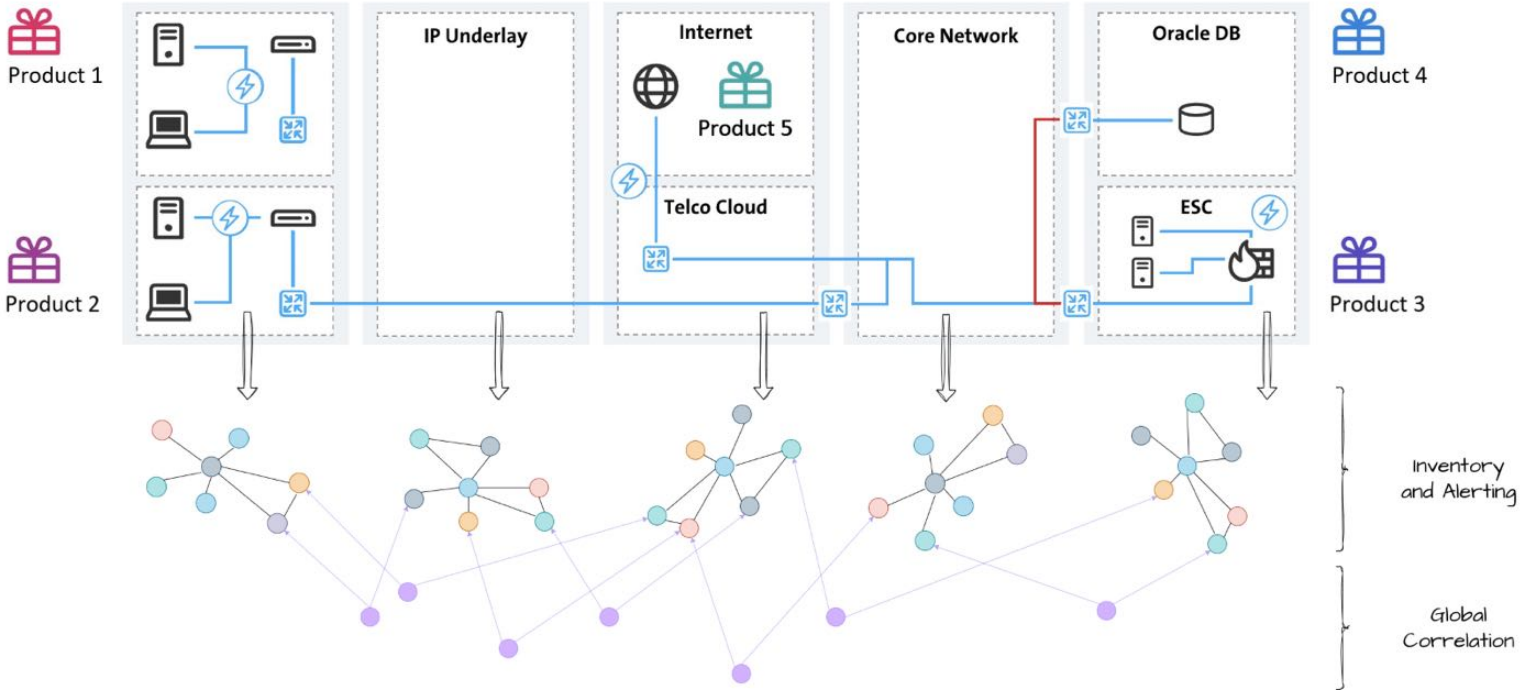
Large-scale graph processing

- Collaboration between USI and Swisscom
  - ✦ Swisscom is Switzerland's leading telecom operator
  - ✦ Reinventing streaming architectures for real-time data and scalable graph analytics



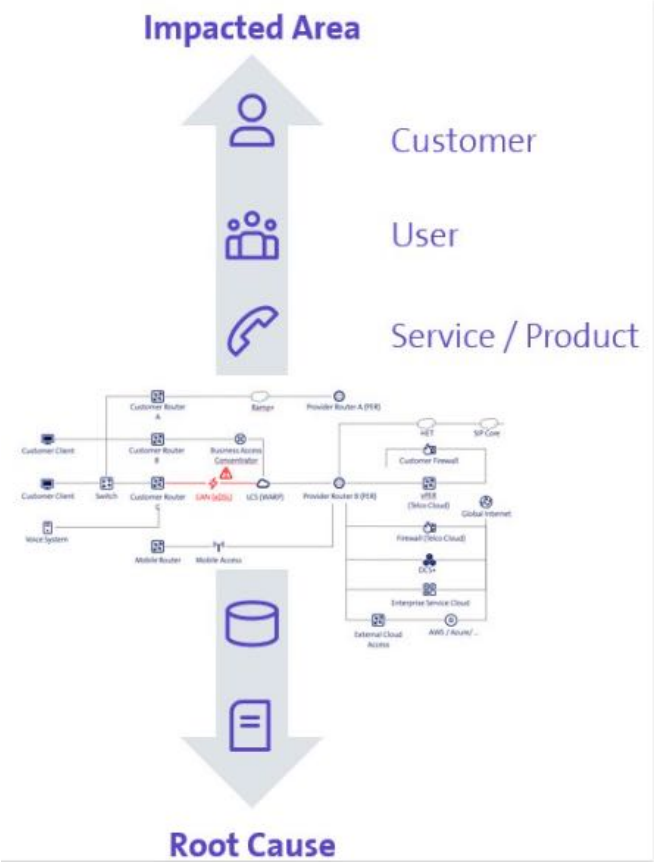
Large-scale graph processing

- The Monty Knowledge Graph
  - ✦ Real-time graph with 200M+ vertices and 20+ data sources



Large-scale graph processing

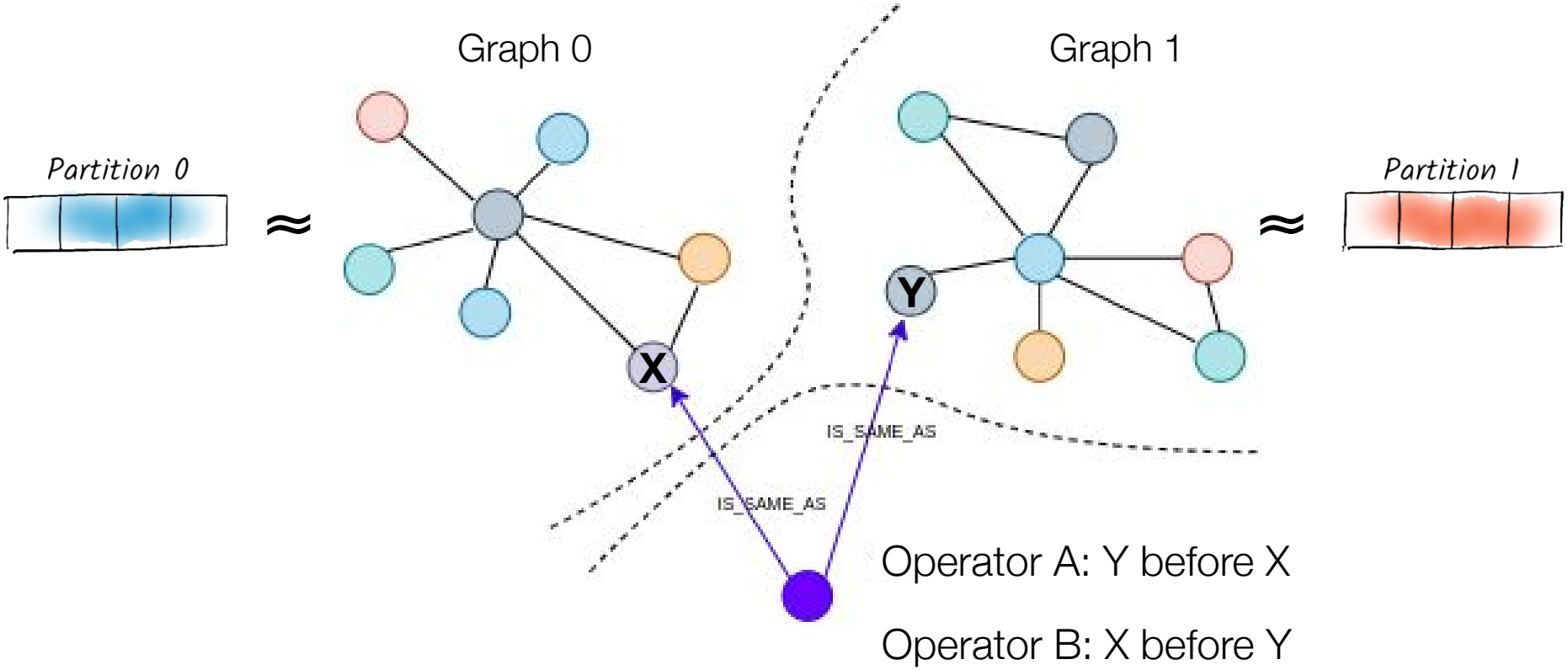
- The Monty Knowledge Graph



Monty calculates **impact** across all systems, by **correlating inventory and event data** to reduce incidents and downtime.

Large-scale graph processing

- Scalable but inconsistent cross-graph updates
  - ✦ Graphs implemented with Kafka partitions

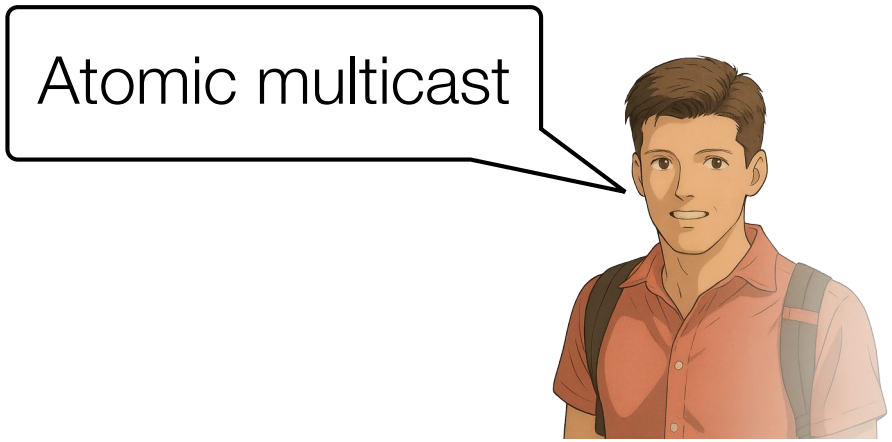


Large-scale graph processing

- Why observing consistent event order is important?
  - ✦ Event X: A network router fails
  - ✦ Event Y: The customer calls support to report an issue
  - ✦ Operator A
    - Customer called (Y), then the router failed (X)
    - They might think the customer is reporting an unrelated issue
  - ✦ Operator B
    - The router failed (X), and then the customer called (Y)
    - They correctly infer the router failure caused the call

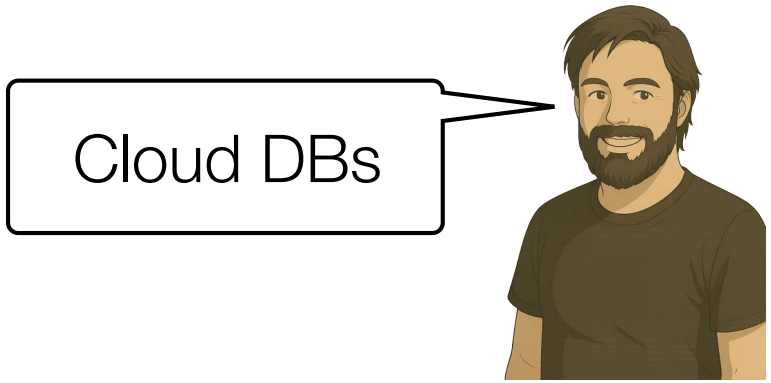
Large-scale graph processing

- How do consistent event order and consensus relate?
  - ✦ Consensus ensures processes agree on the order of events
  - ✦ A special use of consensus called atomic multicast
  - ✦ Atomic multicast keeps Kafka partitions consistent



Database replication

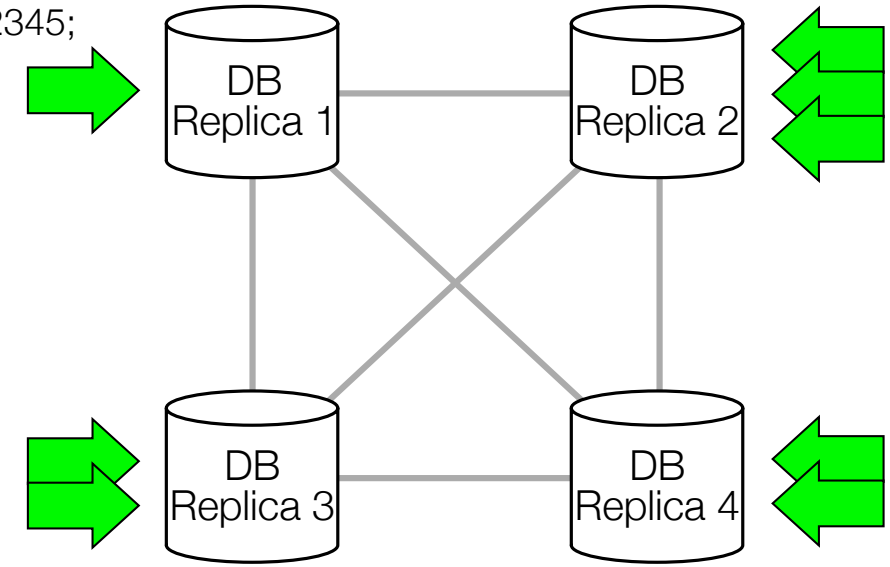
- Replication for fault tolerance and performance
  - ✦ Failure of a few replicas doesn't bring the service down
  - ✦ Every replica contributes to the execution of transactions
- From research prototypes to production
  - ✦ Galera Cluster for MySQL
  - ✦ MySQL Group Replication



Database replication

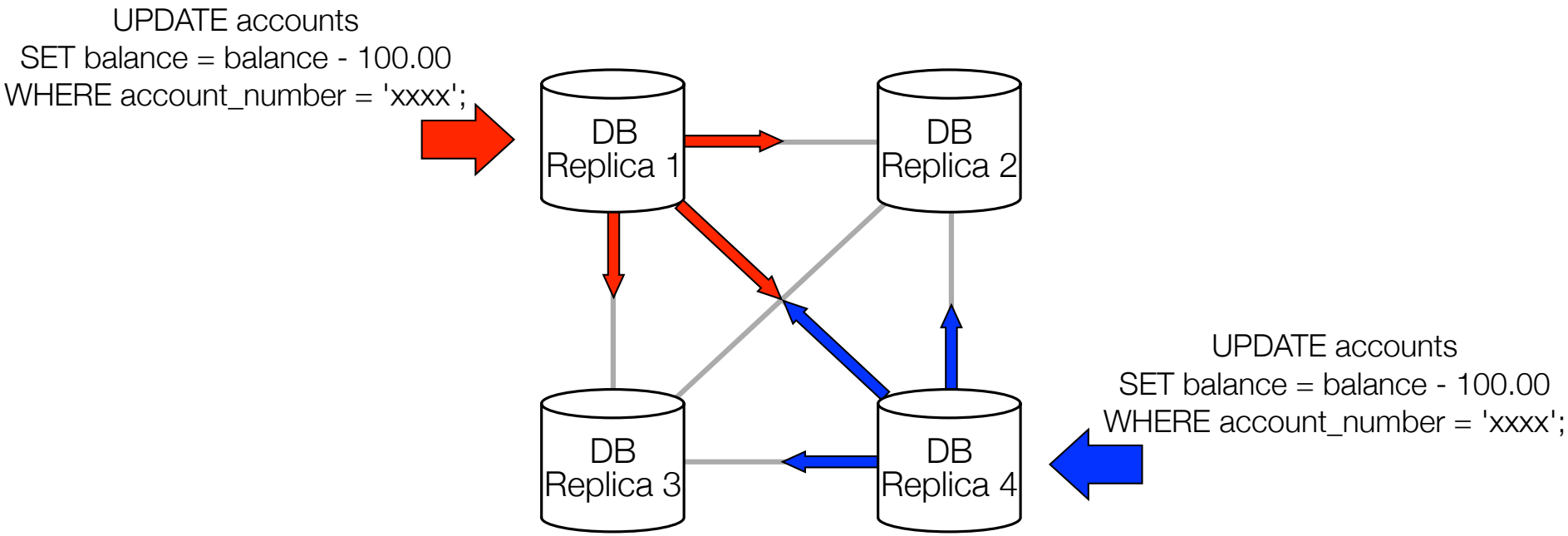
- Deferred Update Replication (with read-only transactions)

SELECT account_number, balance
FROM accounts WHERE c_id = 12345;



Database replication

- Deferred Update Replication (with update transactions)



Database replication

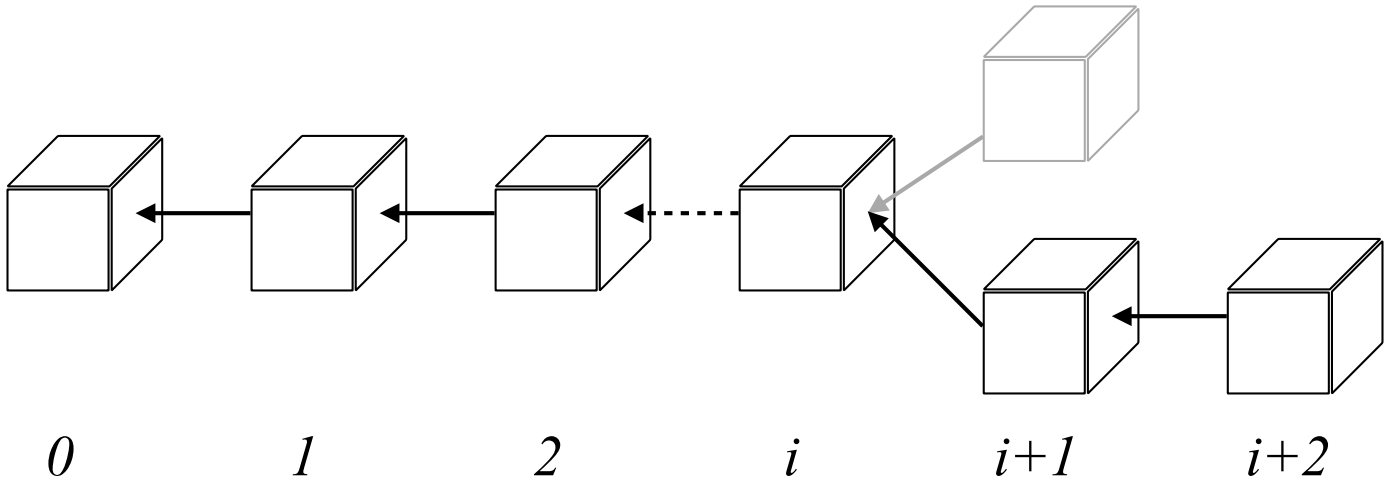
- Deferred Update Replication (with update transactions)
  - ✦ Update transactions
    1. tentatively executed
    2. certified, and
    3. possibly committed
  - ✦ Consensus orders update transactions and ensures that all servers agree on the order transactions are certified

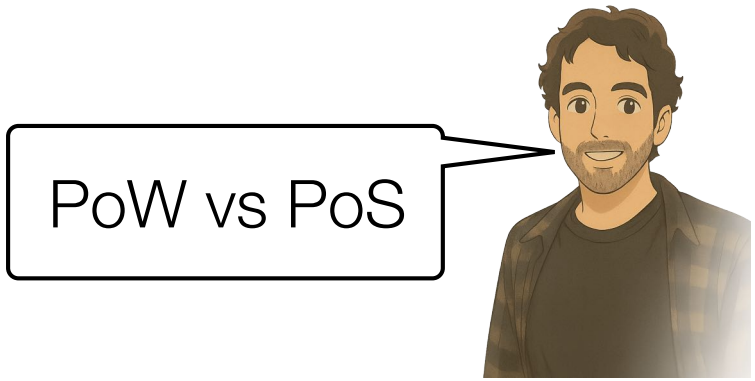
Database replication

- Deferred Update Replication (with update transactions)
  - UPDATE accounts SET balance = balance - 100.00 WHERE account_number = ‘xxxx';
  - ➡ ordered before ➡ : ➡ commits and ➡ aborts
  - ➡ ordered before ➡ : ➡ commits and ➡ aborts

Blockchain

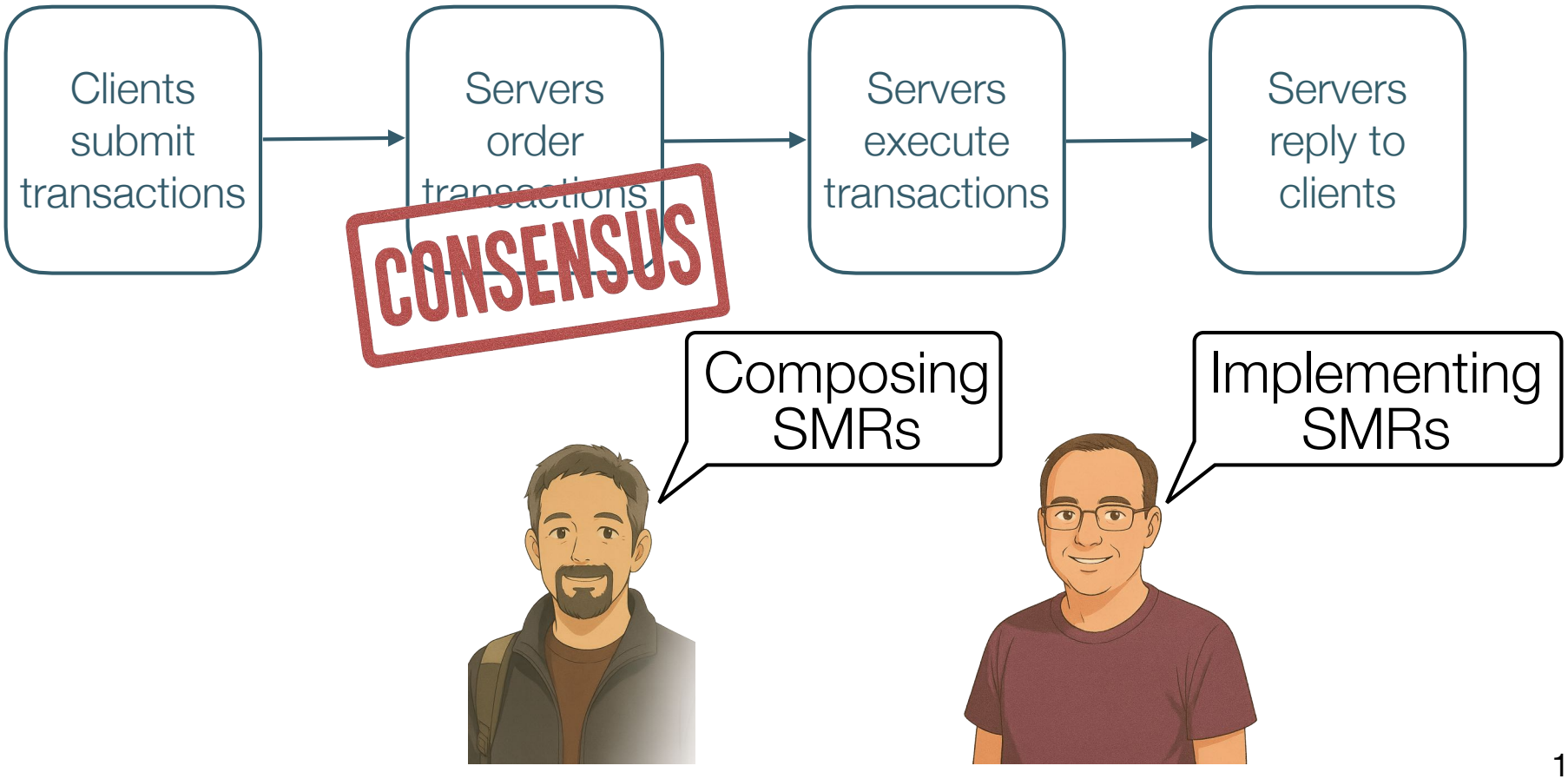
- An immutable ledger, an append-only log of transactions





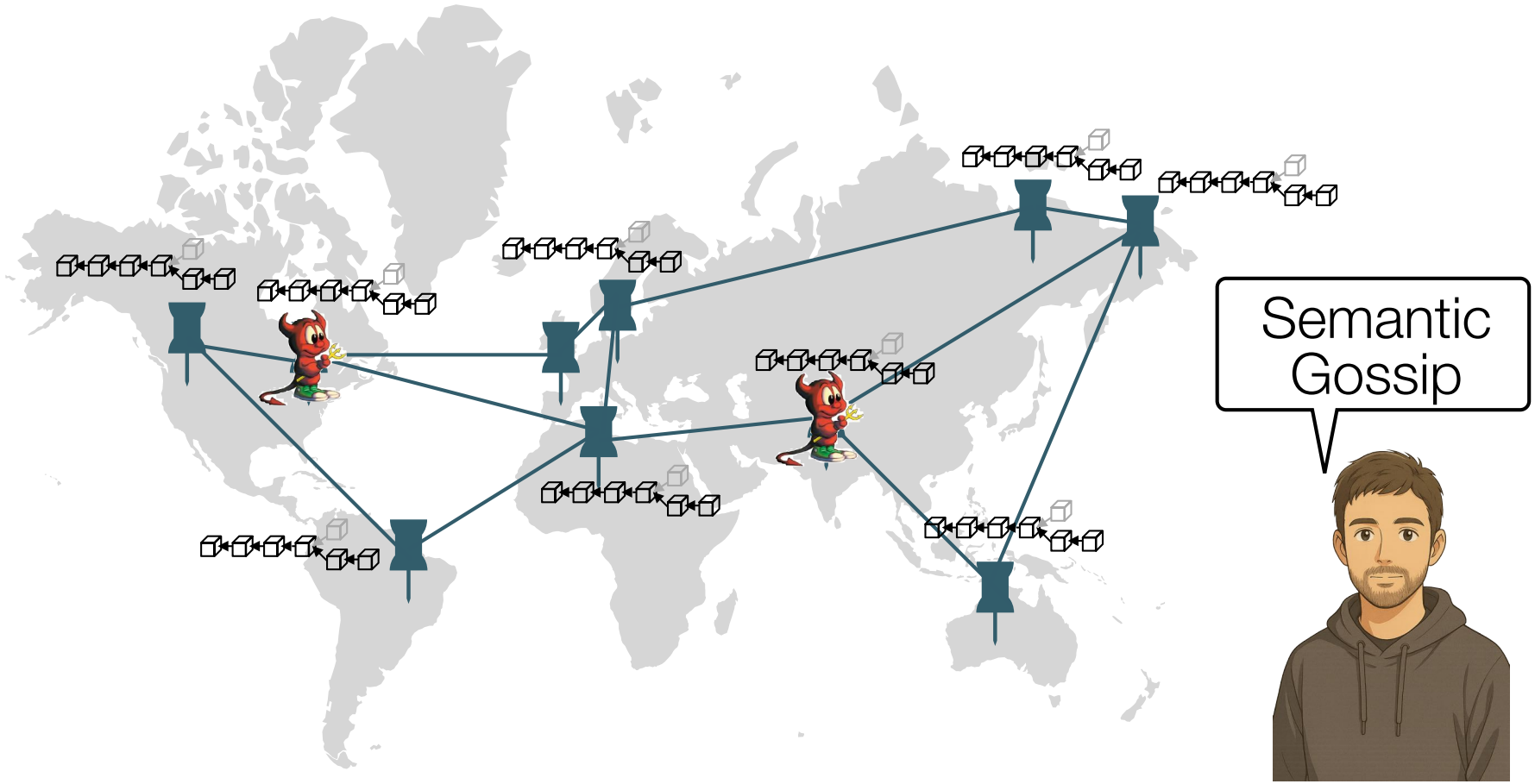
Blockchain

- State machine replication (SMR)



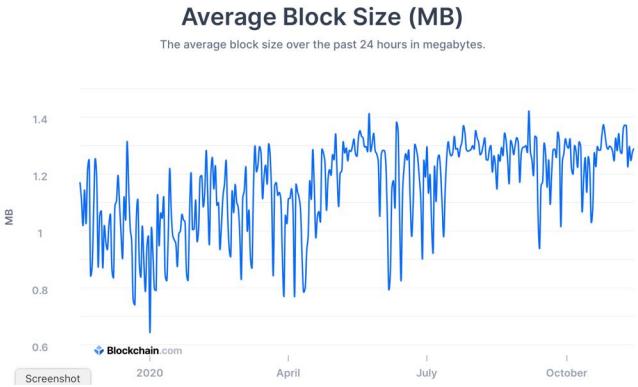
Blockchain

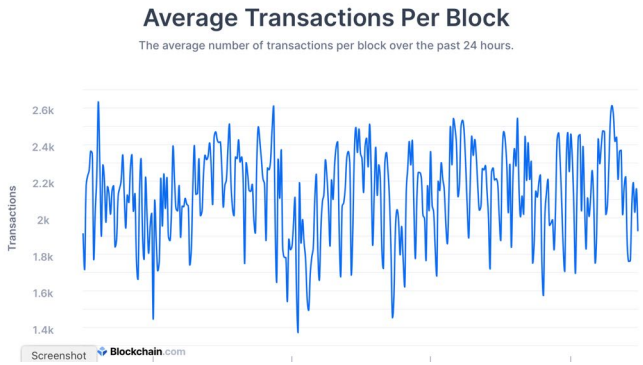
- Network of untrusted geographically distributed parties

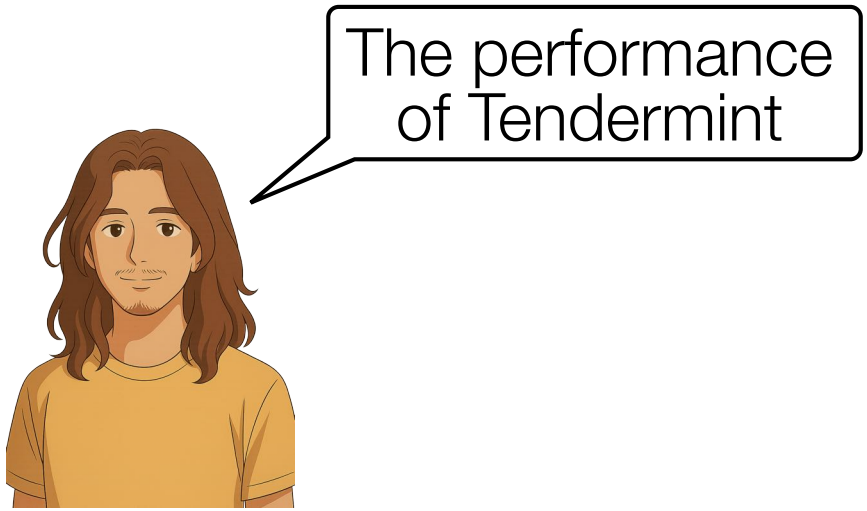


Blockchain

- Understanding blockchain performance





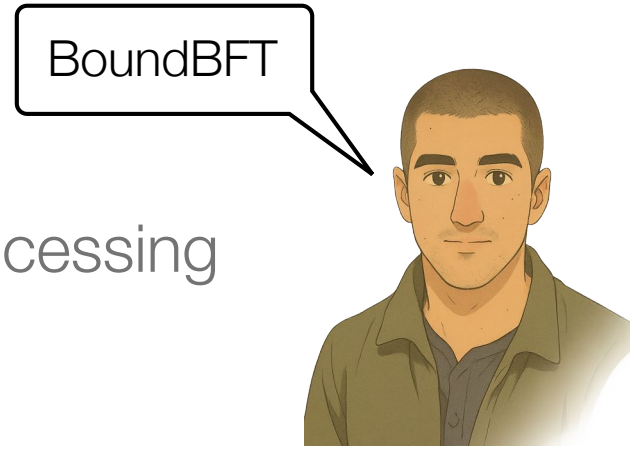


The future

- Novelty driven by…
  - ✦ New use cases and applications
  - ✦ New hardware environments and insights
  - ✦ Taming the complexity of correctness
  - ✦ Understanding performance

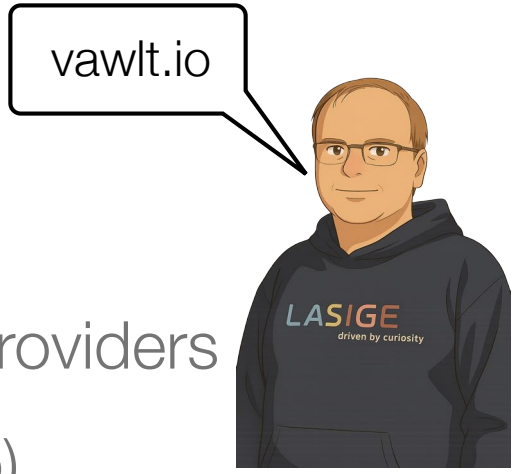
The future

- Synchronous BFT Consensus
  - ✦ Strict bounds for communication and processing
  - ✦ Tolerates more malicious servers
- Replication state management
  - ✦ How quickly replicas can catch up
  - ✦ Optimizing for data structures

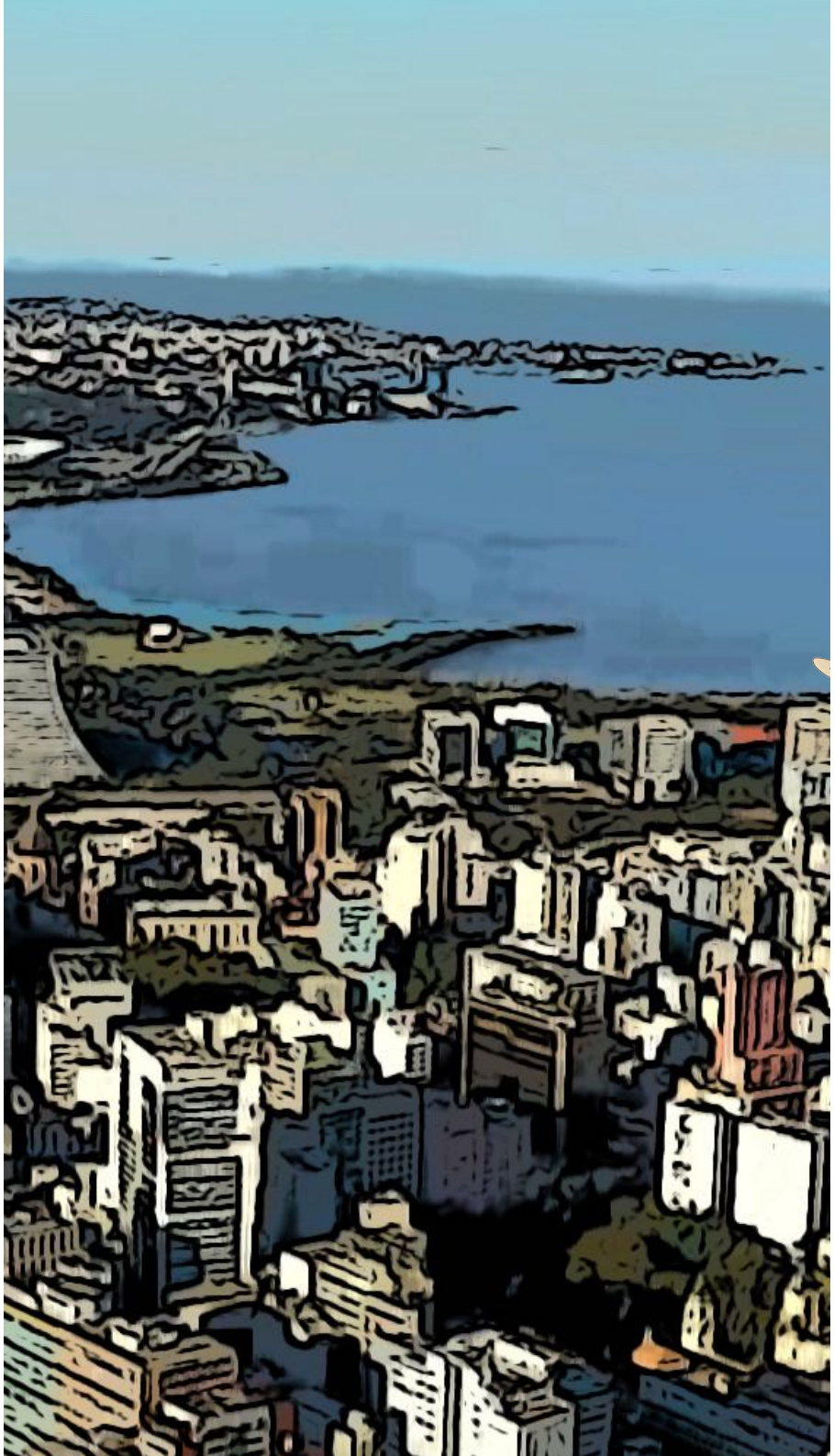


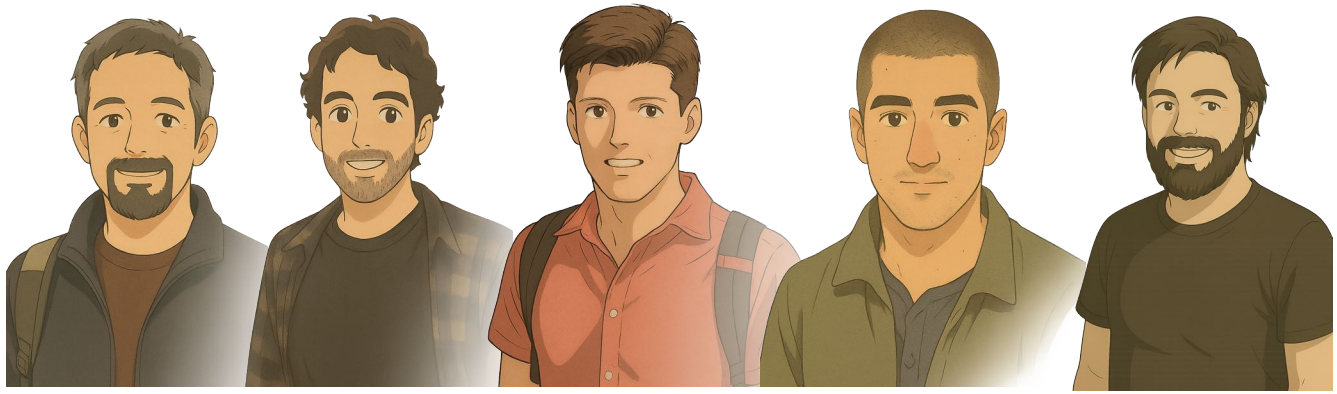
The future

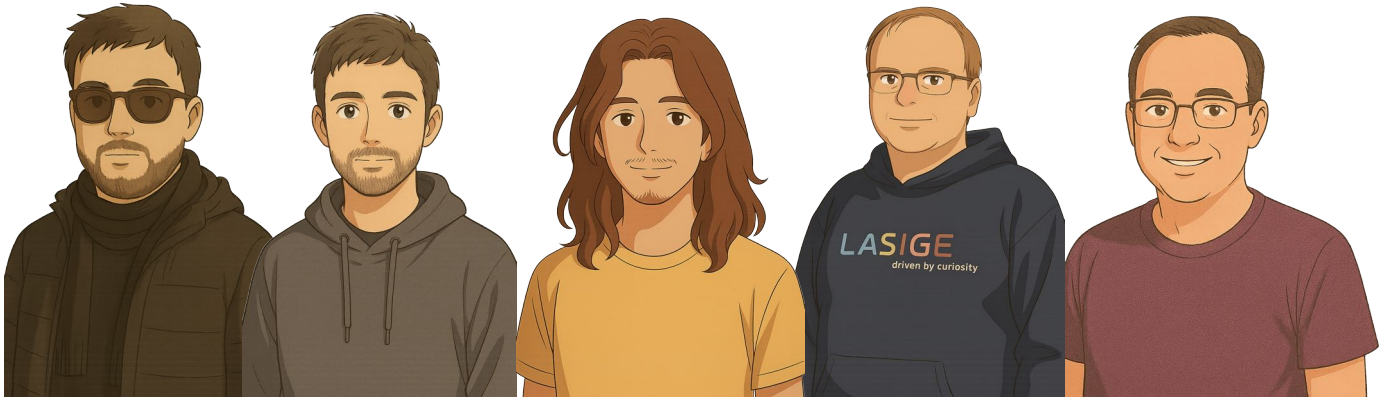
- From research to production
  - ✦ Cloud storage challenges and motivation
  - ✦ Fault-tolerant approach using multiple cloud providers
  - ✦ Development of a commercial service (vawlt.io)



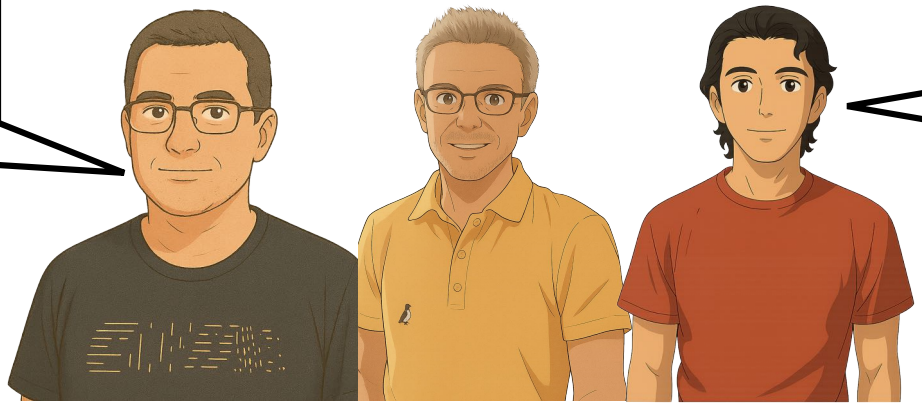
http://www.inf.usi.ch/faculty/pedone/







100 years of collaboration



Uma carreira consistente