DATA SAFETY IN THE CLOUD

An Overview

INTRODUCTION: WHAT IS THIS WORK ABOUT?

- Businesses and individuals put growing amounts of data on the cloud
- Cloud providers consolidate storage and computations
  -> Easier management
  -> Lower costs
- Providers use commodity hardware to bring down costs

Google Cloud Platform

This Work:

- Looks at product lineups of different providers
- Looks at data safety mechanisms of providers

Microsoft Azure

KEY CONCEPTS: SAFETY AND SECURITY

Security:

- Protection from malicious attackers
- Make information inaccessible for unauthorized parties

Safety:

- Protection from hardware failure
- Protection from transmission errors

➡ This work‘s topic: Data Safety

KEY CONCEPTS: DURABILITY AND AVAILABILITY

Availability:

- Percentage of time the system is online
- Typical values: 99% to 99.99%

Durability:

- Metric for data safety
- Percentage of objects retained over the period of one year
- Typical values: ~99.999999999% (11-nines)

KEY CONCEPTS: BACKUP AND REDUNDANCY

Backup:

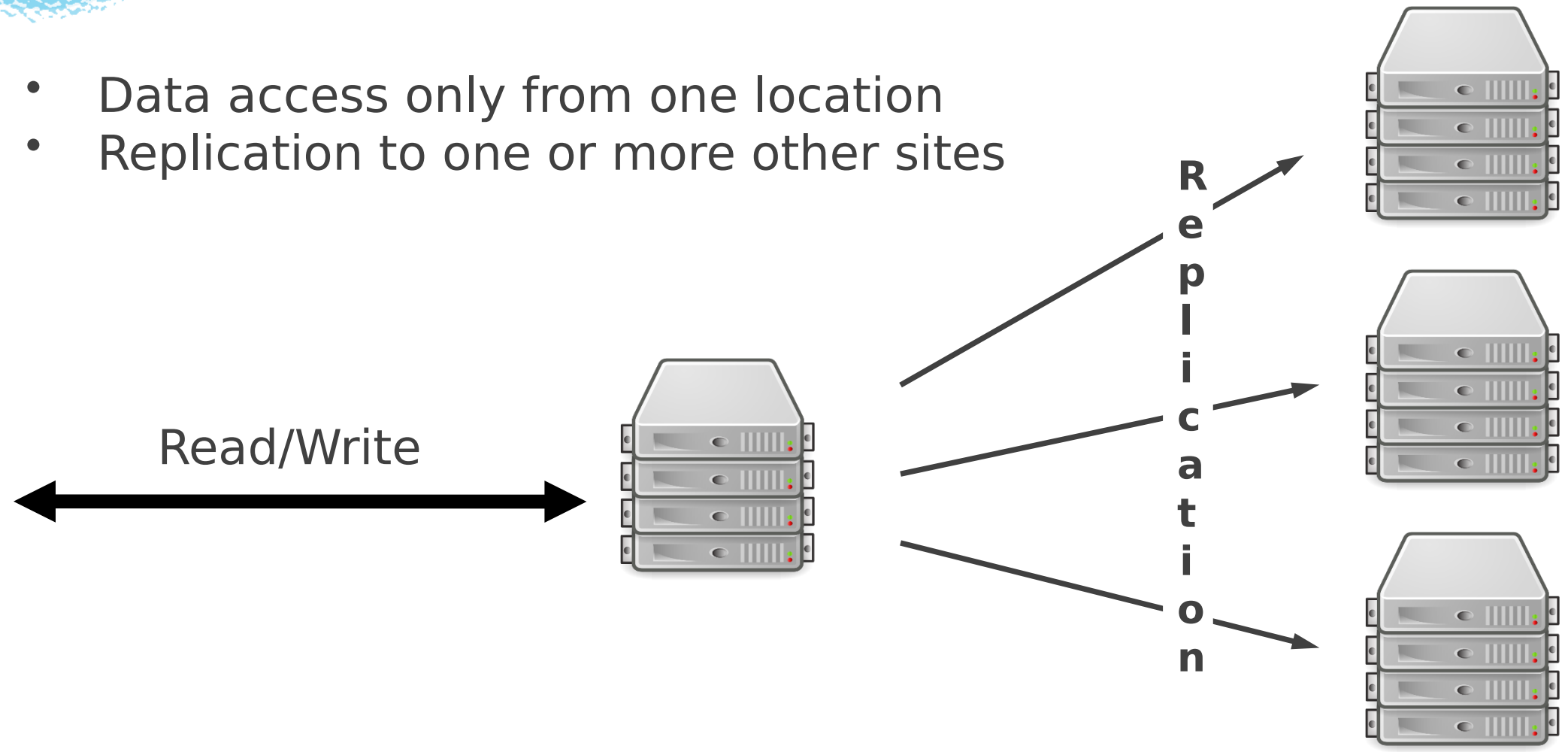
- Point-in-time copy of data
- Used for recovery after catastrophic failure or deletion

Redundancy:

- Copy of data continually kept up-to-date
- Used as a failover in case of a site failure

KEY CONCEPTS: REDUNDANCY

- Data access only from one location
- Replication to one or more other sites



PRODUCT LINEUP: COLD OBJECT STORAGE

Usages:

- Archiving
- Disaster Recovery
- Backups

Technologies:

- Redundancy (local/zone/global)

Products:

- Google Cloud Nearline
- Google Cloud Coldline
- Amazon S3 Infrequent Access
- Amazon Glacier
- Microsoft Backup

PRODUCT LINEUP:
DATA WAREHOUSE & BIG DATA

Usages:

- Business Analysis
- Process Optimization

Technologies:

- Automatic Backups
- Local Redundancy

Products:

- Google Bigtable
- Microsoft SQL Data Warehouse
- Amazon Redshift
- Microsoft Data Lake

PRODUCT LINEUP: GLOBALLY SCALING DATABASE

Usages:

- Globally Distributed Applications
- IoT
- Mobile

Products:

- Google Cloud Spanner
- Microsoft Cosmos DB

Technologies:

- Automatic Backups
- Geo Redundancy

PRODUCT LINEUP: NOSQL DATABASE

Usages:

- Key-Value Storage
- Low Latency

Technologies:

- Manual Backups
- Local or Geo-Redundancy

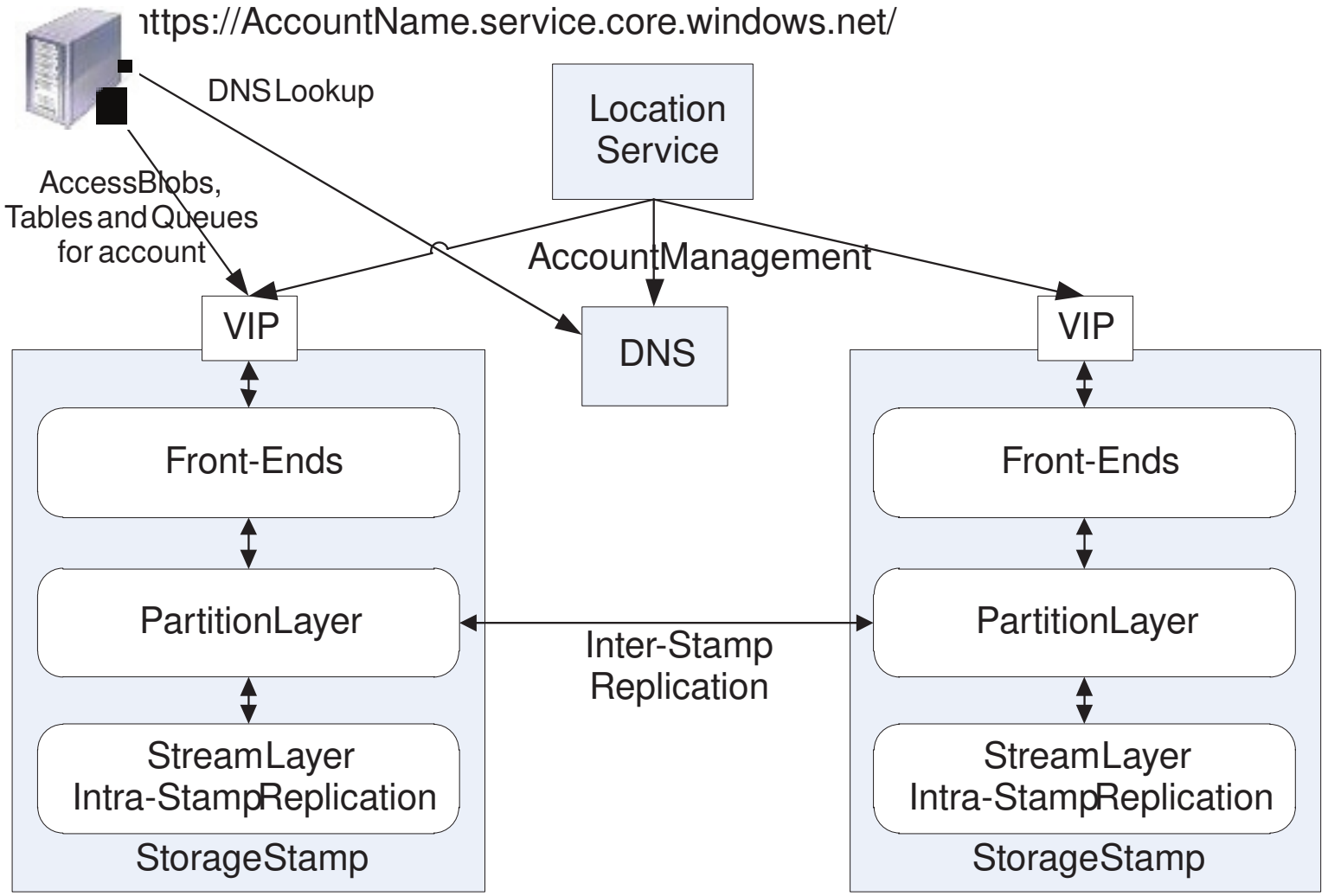
Products:

- Microsoft Table Storage
- Google Cloud Datastore
- Amazon DynamoDB

TECHNOLOGY: MICROSOFT AZURE STORAGE

Layered architecture with redundancies at every level

Following: Top-down overview of the system



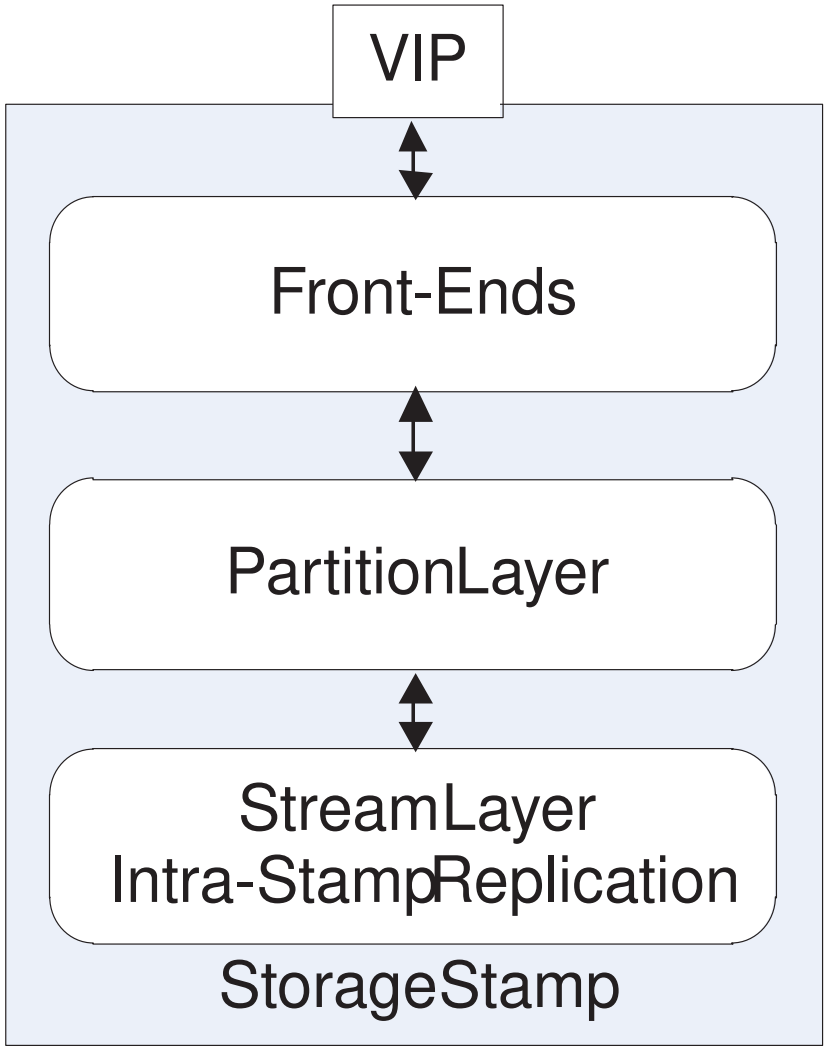
TECHNOLOGY: MICROSOFT AZURE STORAGE

VIP (Virtual IP)
- Entry-point to the stamp

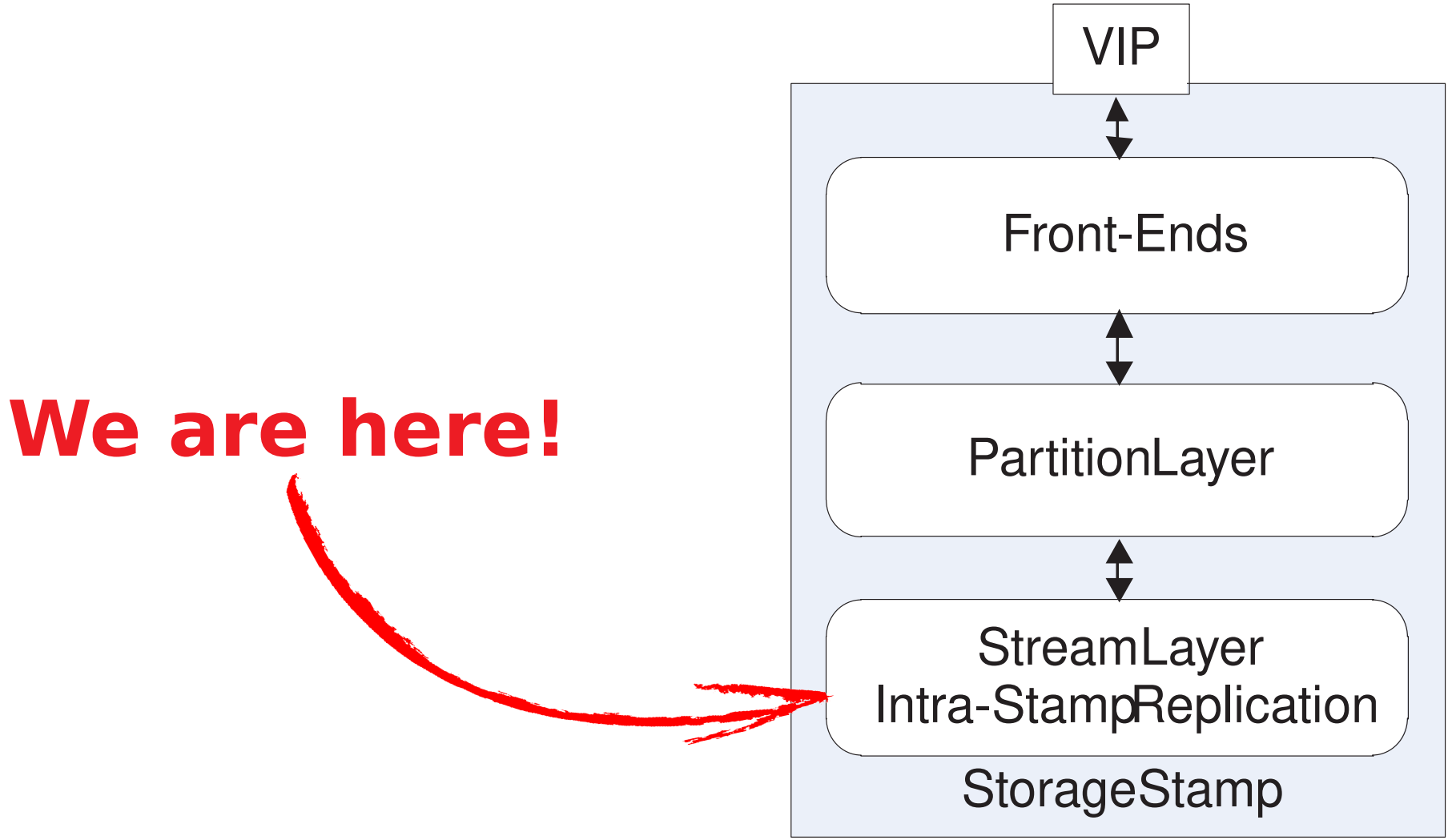
Front-Ends
- Authentication & Authorization
- Routing to Partition Server

Partition Layer
- Abstractions to data in Stream Layer
- Object caching
- Object-based inter-stamp replication

Stream Layer
- Data storage
- Intra-stamp Replication



TECHNOLOGY: MICROSOFT AZURE STORAGE



TECHNOLOGY: MICROSOFT AZURE STORAGE

Streams

Stream

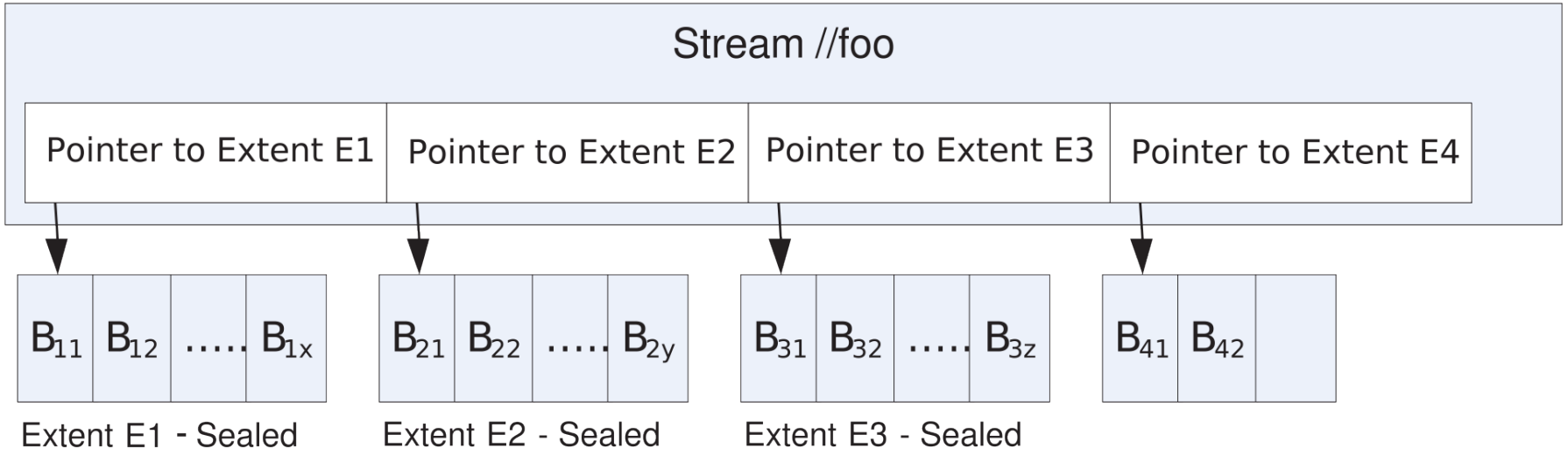
- Looks like traditional file to Partition Layer
- Internally: Only pointers to extents
- Append-only

Block

- Smallest unit of data
- Variable size

Extent

- Sequence of Blocks
- Append-only
- Target size: 1GB
- When full: Sealed



IMPLICATIONS: WHAT TO CHOOSE

The ideal solution for your program depends on its requirements:

Need safe storage?
→ Choose any solution with high levels of redundancy
Need to process lots of data?
→ Choose any big data or data warehouse solution
Need NoSQL storage?
→ Choose any NoSQL solution (but make regular backups)
Need global, consistent and fast access to your data?
→ Choose any globally spanning database
Need multiple things?
→ Combine your solutions e.g. have a safe storage for important data but copy it to a big data solution for processing

IMPLICATIONS: HOW TO DESIGN

General design guidelines:

- Do not rely on locally mounted storage!
  - Block storage is available but inflexible
  - Use more flexible and safer storage options
- Seperate storage functionality from the rest of the program
  - Makes it easier to migrate data to a different provider later
  - Makes it easier to adapt to provider API changes
- Keep applications stateless
  - Easily update application if provider or storage needs change

CONCLUSION: SUMMING IT ALL UP

- Providers offer products and services for almost every use-case
- Providers invest heavily into their architectures and the technologies behind them
- Data safety is achieved through several technologies:
  - Erasure Coding & RAID
  - Data Redundancy across one or more facilities
  - Automatic Backups
- The challenges providers are facing today are about maintaining strong consistency and high performance across a global database

Thank You For Listening