Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

Spilgames Storage Platform

Enrique Paz

Senior Backend Developer

21/03/2013

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

Introduction

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

About Me

- Passionate Erlang developer
- Testing enthusiast
- Love beautiful code!

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

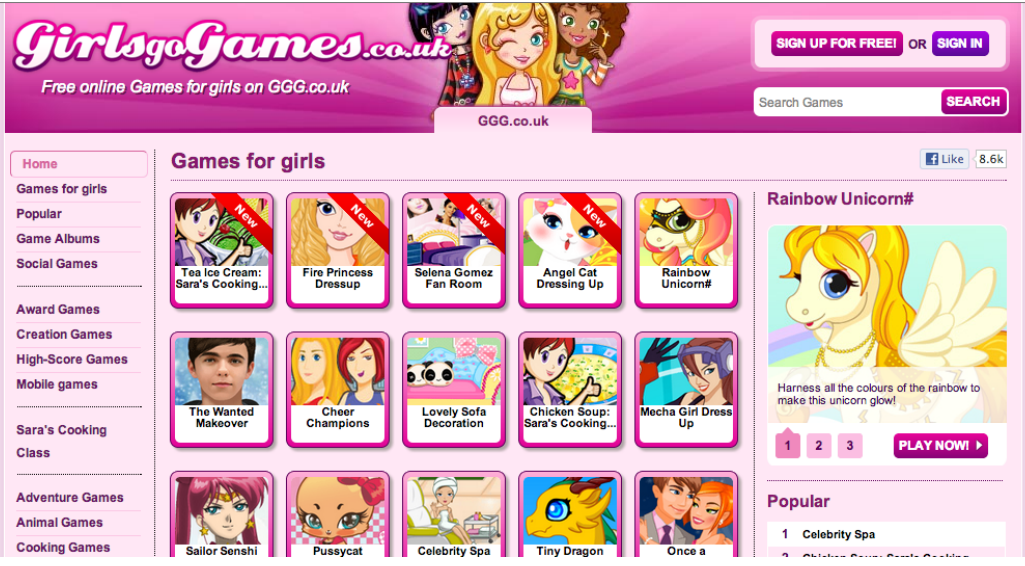
Spilgames

- Gaming Platform
- Serving data to 190+ countries world-wide
- 200+ million unique users per month
- Multiple Platforms: Desktop, Mobile Native & Web
- 300+ employees
- Offices in The Netherlands & China
- Revenue: Advertising & EUM

Gaming Portals

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions



Games

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

Another Storage

Another Storage

Introduction
- Spilgames & Me

Another Storage
- Motivation

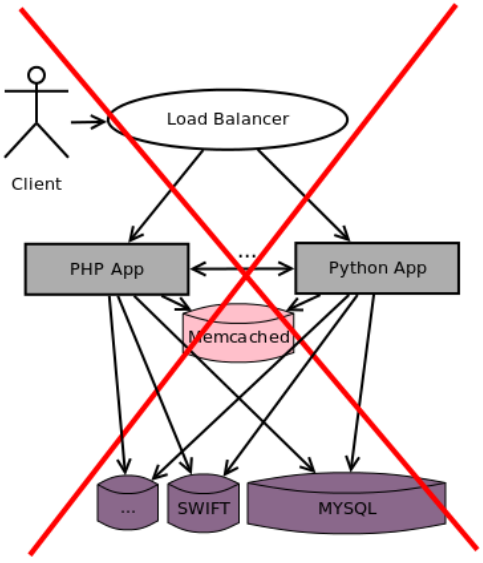
System Properties
- Overview
- Internals
- Versions & Shards

Worldwide
- Scaling Datacenters
- Disaster Scenarios

Lessons Learned
- Currents Status
- Contributions

LAMP Stack & MySql

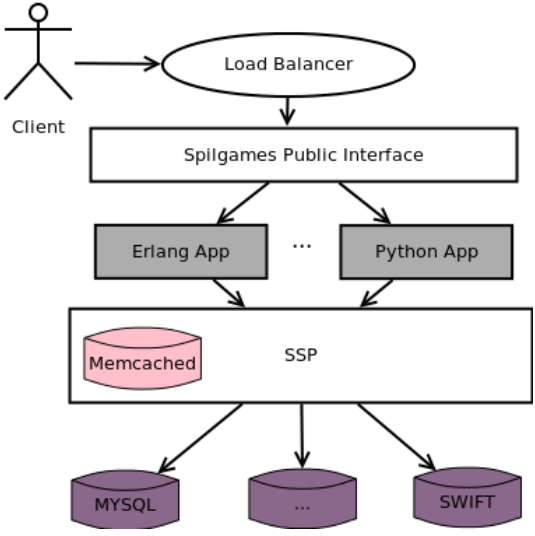
- Not all developers are DB experts
- Difficult to shard the databases
- Storage model all over the place
- Security
- Performance
- Caching
- . . .



Our Ambition

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

- Transparent sharding layer
- Sharding on data ownership
- High availability
- Centralized caching layer
- Storage engine agnostic
- One strict data model
- Transparent storage changes
- Scaling geographically



Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

System Properties

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

Mindset

- Be always available
- Avoid global locks
- Accept change as the only constant
- Embrace inconsistencies
  - Hardware breaks down (power failures)
  - Version mismatches (upgrading system not atomic)
  - State mismatches (adding new machines)

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

A Key-Value Store With Schema

- **Buckets**
  - Largely generated OTP applications
  - Offer a CRUD-like interface (with filters)
- **GID**s (64 bits) identify the data owner
  - user
  - game
- Buckets can use different **storage engines**
  - Several MySql tables in different databases
  - Just a binary storage (SWIFT)
  - ...
- Data for a bucket/GID is cached
- Requests can be atomic per bucket/GID

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

Optimistic Operations

- Speed > Consistency
- Losing some updates in case of crash is affordable
- Act first on cache and then on disk
- No warranties of eventual consistency upon crashes
  - i.e. Activity feeds
  - i.e. Popular games list

Pessimistic Operations

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

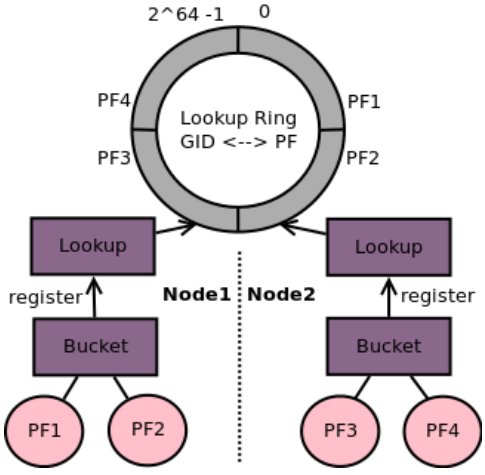
- Consistency is key and confirmation is required
- Dealing with critical data
- Persist data and, upon success, update cache
  - i.e. Payments
  - i.e. Personal information

How does it work?

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

System Components

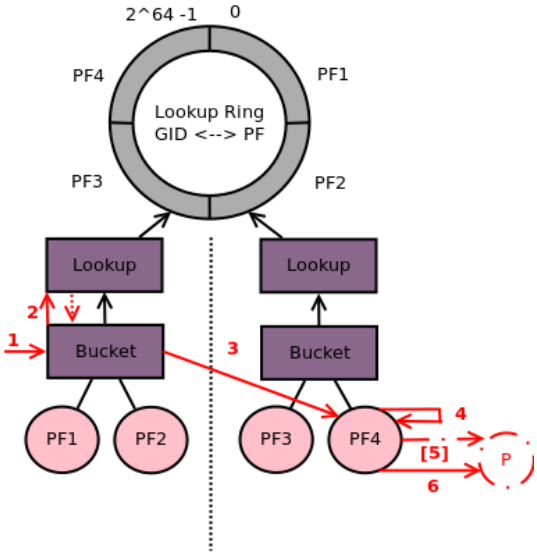
- **lookup** application in all nodes. Uses a **hashing ring** (mnesia):
  - replicated in all nodes
  - ram_copies
  - dirty reads
  - transactional writes
- **Buckets** have Pipeline Factories
- Buckets register **PF**s in lookup



Tracing Pessimistic Operations

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

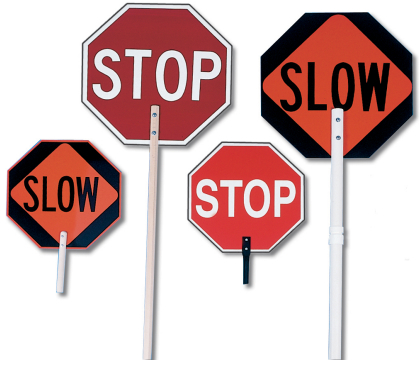
1. Bucket/GID request in a node
2. Local lookup to find a PF
3. PF receives request
4. PF builds job
5. PF ensures Pipeline for GID
6. PF queues operation in Pipeline



Wait a minute. . .

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

- “Why do we need pipelines?”
- “Sequential == Bottleneck !!!”
- “Don’t you guys know Erlang is about **parallelizing** work?”



About Pipelines

- **CONS**
  - Sequential (read) access for hotspots
    i.e. Popular games
  - Optimization: read from SSP cache in PF

About Pipelines

- **CONS**
  - Sequential (read) access for hotspots
    i.e. Popular games
  - Optimization: read from SSP cache in PF
- **PROS**
  - No need for storage engines to support global locks
  - A bucket can combine several engines

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

About Pipelines

- **CONS**
  - Sequential (read) access for hotspots
    i.e. Popular games
  - Optimization: read from SSP cache in PF
- **PROS**
  - No need for storage engines to support global locks
  - A bucket can combine several engines
- Requests to most GIDs (users) are evenly distributed

Schema Versions

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

- Schema Versions determine allowed operations and storage(s)
- Client is not aware of them
- Max 2 schema versions of a bucket at the time
- Schema version can be changed at bucket/GID level

Shards

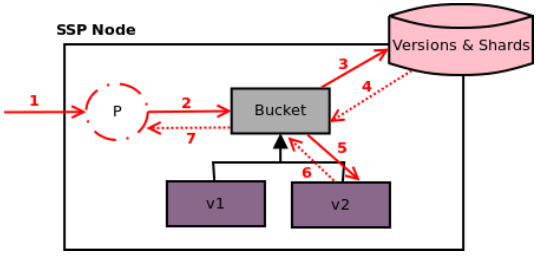
Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

- Useful for partitioning big blocks of data
- Shard points to the physical location of the data
- Sharding rules are bucket specific. Default is GID % Shards
- bucket/GID combinations can be migrated between shards

Working With Versions & Shards

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

1. *insert(Bucket, Gid, Value)*
2. *insert(Gid, Value)*
3. *get_vs(Bucket, Gid)*
4. *{v2, Shard1}*
5. *build_job(insert, Gid, Shard1)*
6. *{ok, InsertJob}*
7. *{ok, InsertJob}*



One API To Rule Them All

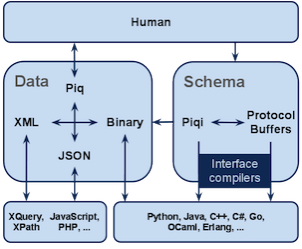
Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

- “Don’t care where it is, just want my data!!!”

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

One API To Rule Them All

- “Don’t care where it is, just want my data!!!”
- PIQI helps with the API



- Erlang client + Protocol Buffers
- HTTP + JSON
- HTTP + Protocol Buffers

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

Worldwide

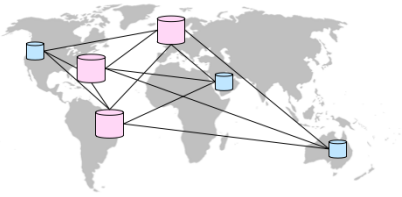
Worlwide

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

spilgames

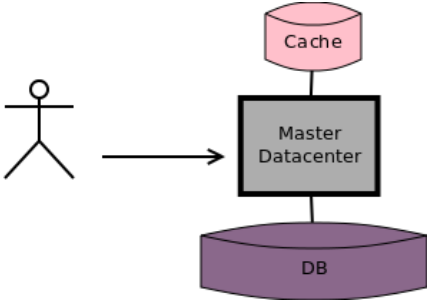
Masters & Satellites

- Master Datacenters
  - Have persistent storage
  - Can own GIDs
  - GIDs can be migrated between Master DCs
- Satellite Datacenters
  - Don't have persistent storage
  - Easy to setup and decommission
  - Virtual/Cloud-based



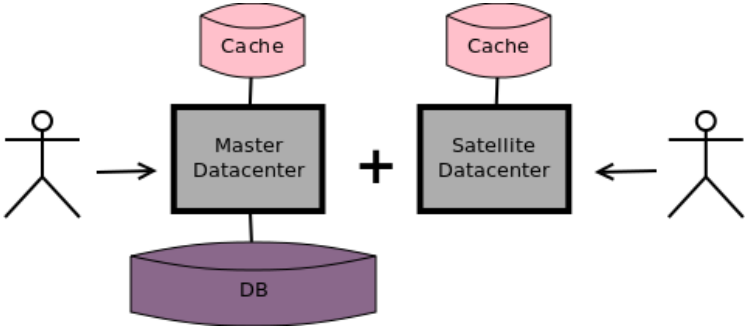
Start With One DC

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions



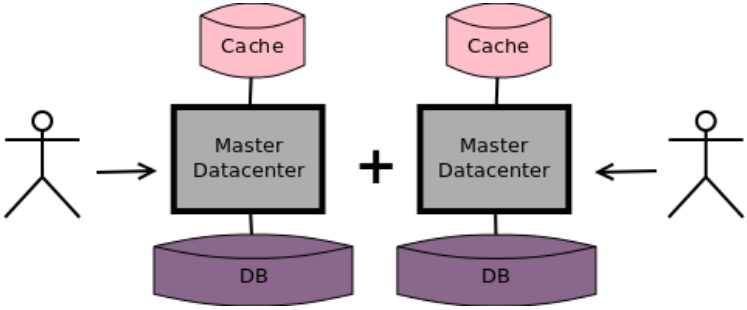
Scale Up A Satellite Where Needed

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions



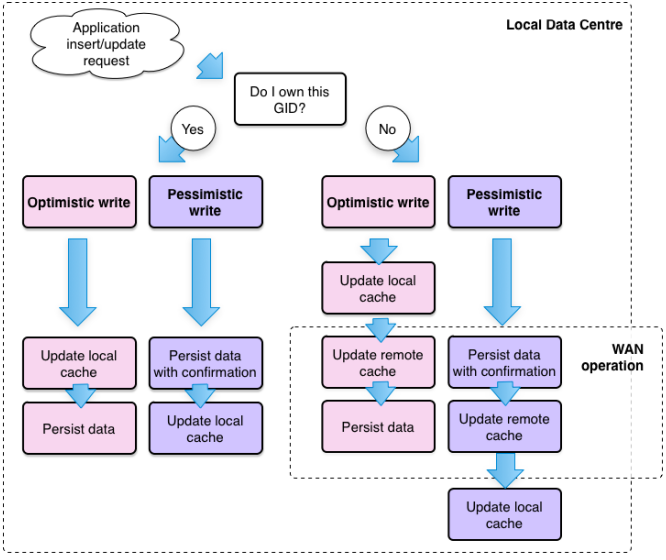
Turn Satellites Into Masters When Ready

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions



Working With Multiple Datacenters

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions



Disaster Scenarios

Introduction
- Spilgames & Me

Another Storage
- Motivation

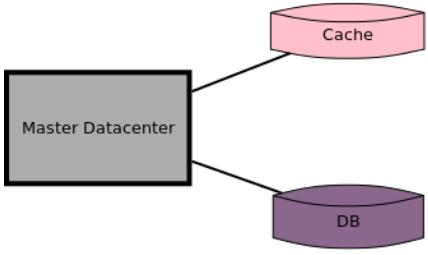
System Properties
- Overview
- Internals
- Versions & Shards

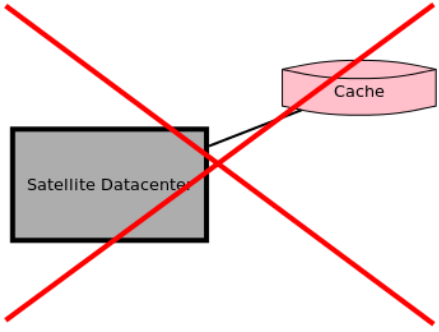
Worldwide
- Scaling Datacenters
- Disaster Scenarios

Lessons Learned
- Currents Status
- Contributions

Losing A Satellite Datacenter

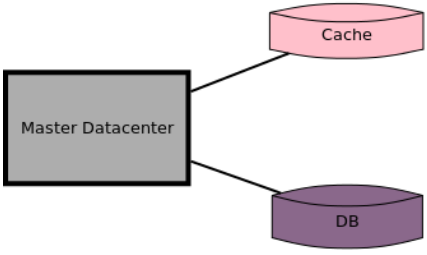
Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

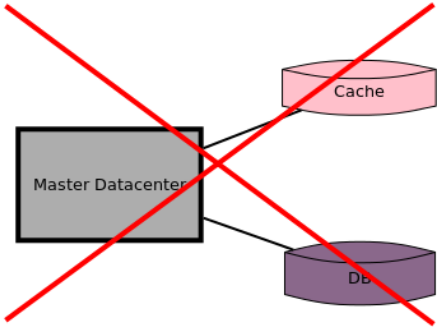




Losing A Master Datacenter

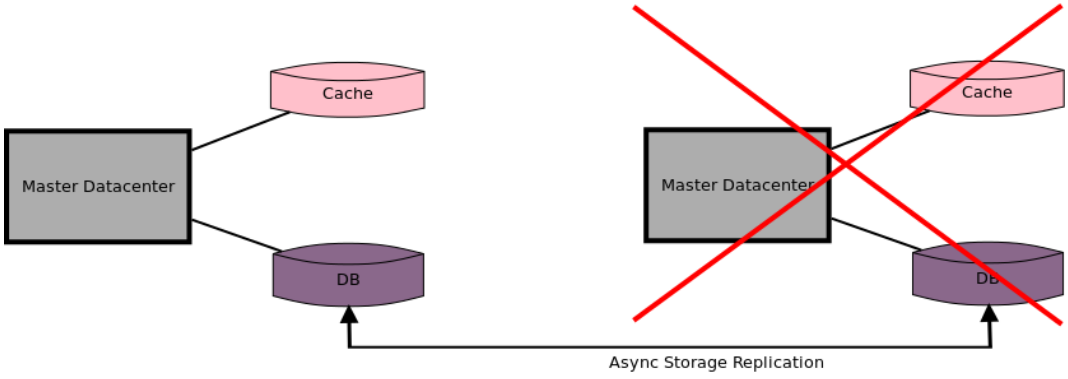
Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions





Losing A Master Datacenter

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions



Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

Lessons Learned

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

Where We Are

- Using simple buckets on LIVE in one DC
- Added relup support for bucket only updates
- Hammering SSP using property based testing
- Integrating restricted search capabilities
- Testing the WAN protocol for inter DC communication
- More buckets to go live in H1 2013
- Satellite DCs coming on H2 2013

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

What We've Used

- Emysql
  - (+) multi-database transaction support
  - (*) multi-timezone support
- Eep0018/Jiffy
- Estatsd
- PropEr
- Poolboy
- Lager
- Rebar
  - (*) semantic versioning, i.e. [">=1.3.1", "<2.0.0"]
  - (*) shared dependencies
  - (*) xref fixes
- Piqi
- BashoBench
  - (+) Several tests on the same plot support

Questions?

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

That's All Folks!

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

Thanks!

Join Us!

Introduction
Spilgames & Me
Another Storage
Motivation
System Properties
Overview
Internals
Versions & Shards
Worldwide
Scaling Datacenters
Disaster Scenarios
Lessons Learned
Currents Status
Contributions

http://www.spilgames.com/careers/job-openings/