Megastore: Solução para as crescentes exigências dos serviços na nuvem

Katharina C. Garcia 317144

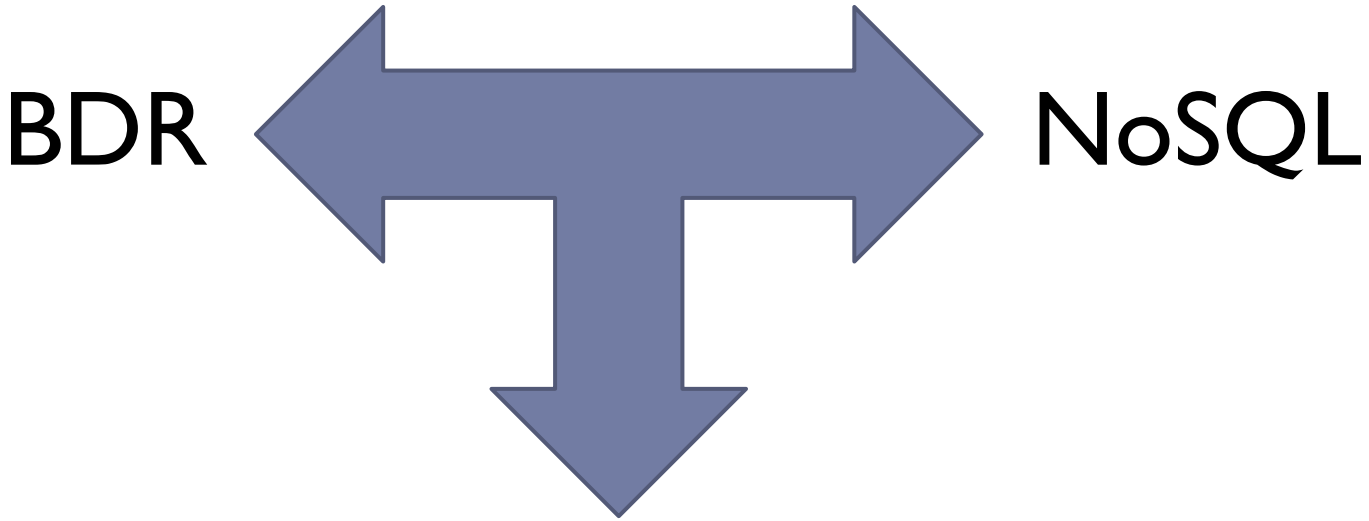
Agenda

- Problema e Solução
- Replicação e Paxos
- Arquitetura
- Particionamento
- Modelo de dados
- Algoritmos
- Tratamentos de falhas
- Análises
- Conclusão

Problema

- Usuários muito exigentes
- Novas aplicações mais poderosas
- Utilização intensiva da Cloud
- Escalabilidade, Disponibilidade, Armazenamento, Consistência, Rapidez

Solução

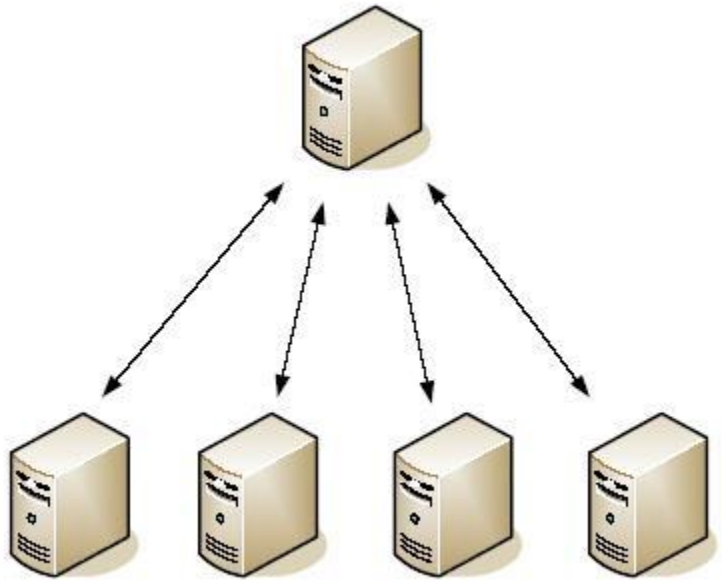


Megastore

- BDR + NoSQL = Alta Disponibilidade + Consistência + Escalabilidade
- Google – App Engine
- Resultado: 3 bilhões de escritas + 10 bilhões de leituras

Replicação

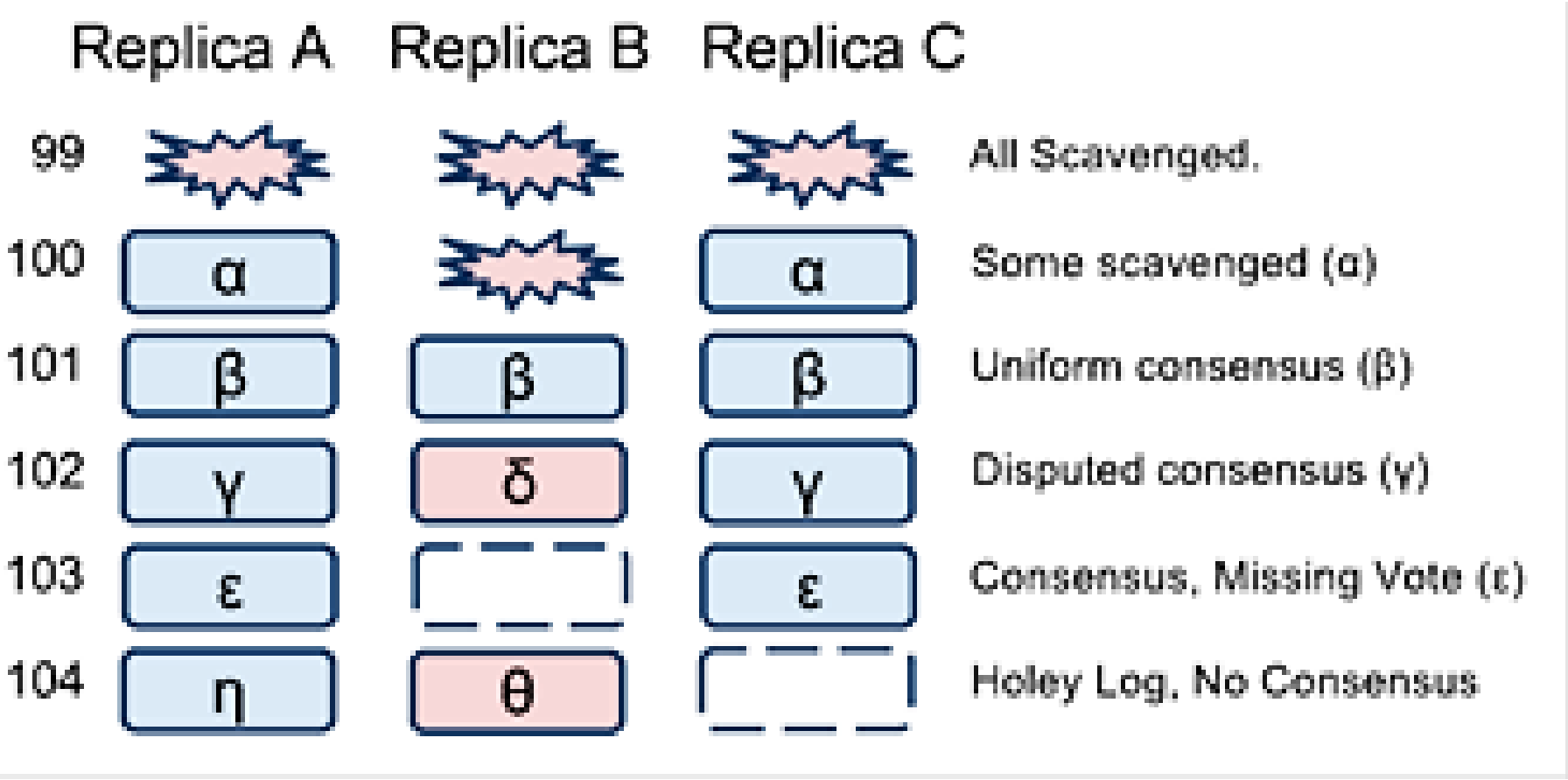
- Replicas distribuídas
- Síncrona
- Visa a: Disponibilidade
- ~~Master/Slave~~



Paxos

- Algoritmo para sistemas distribuídos tolerante a falhas
- Encontra um consenso entre diversas replicas
- Log replicado
- Todos podem iniciar operações
- A maioria escolhe o valor mais apropriado
- Uma instância de paxos para cada posição

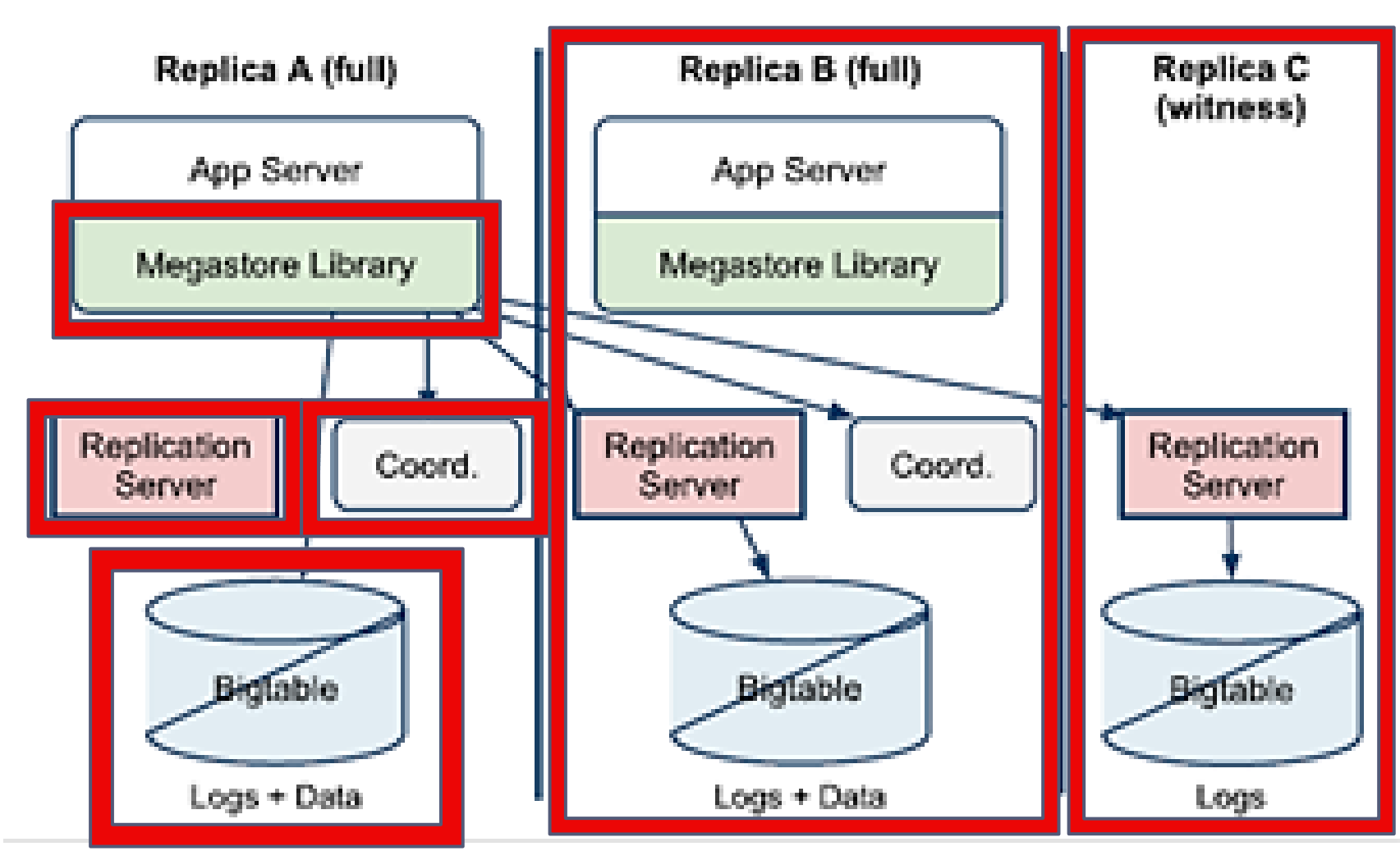
Paxos - Log


Replica A
Replica B
Replica C
99
All Scavenged.
100
α
α
Some scavenged (α)
101
β
β
β
Uniform consensus (β)
102
γ
δ
γ
Disputed consensus (γ)
103
ε
ε
Consensus, Missing Vote (ε)
104
η
θ
Holey Log, No Consensus

Paxos - Melhorias

- Leituras rápidas
  - Líder
  - Coordinator
- Tipos de replicas
  - Full
  - Witness
  - Read-only

Arquitetura



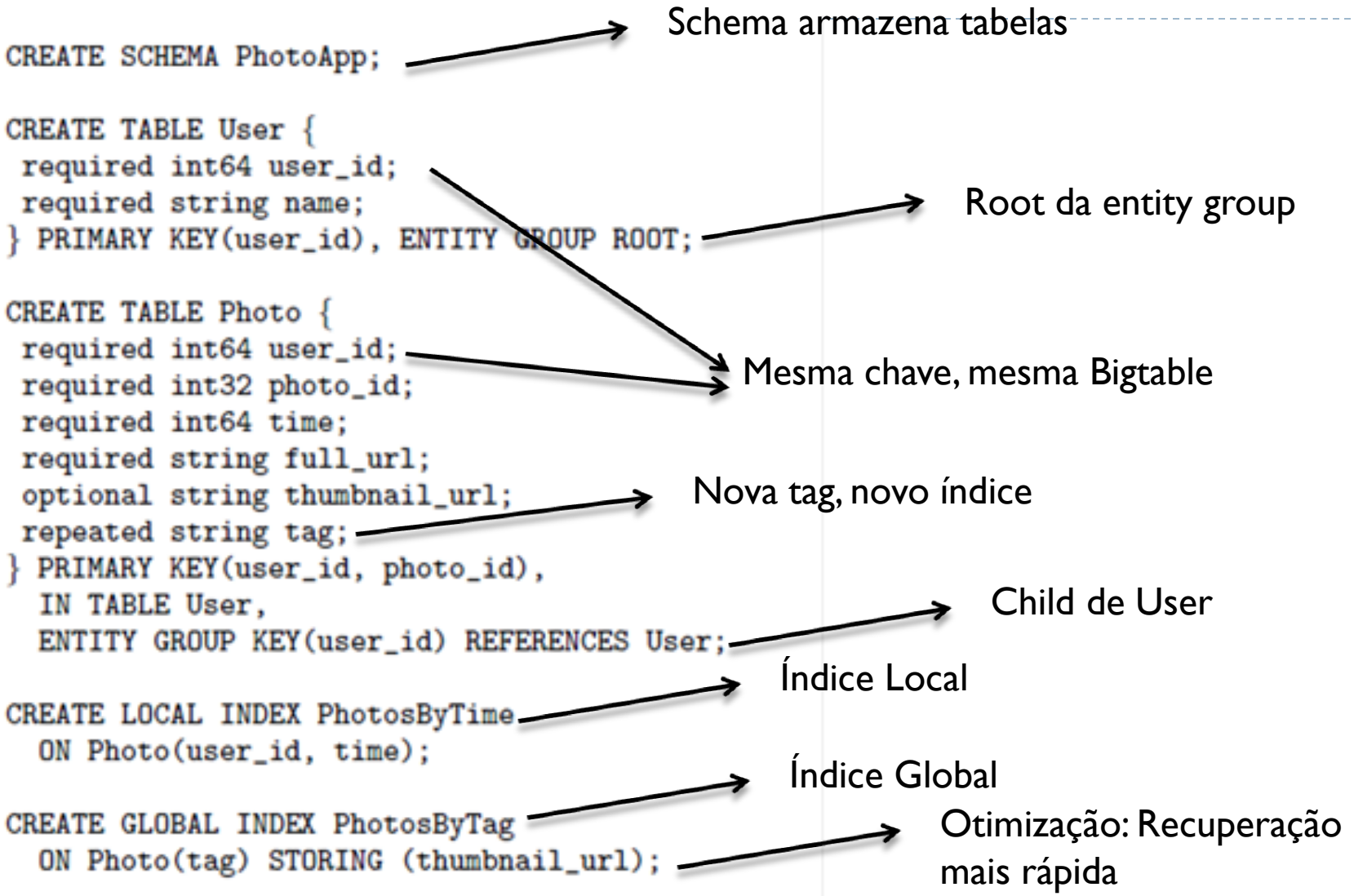
Arquitetura

- Base NoSQL: BigTable
- Instância é escolhida com base da proximidade geográfica
- Dados relacionados e/ou mais acessados = colunas próximas

Particionamento

- *Entity Group*
- Transações ACID internas
- Mensagens Assíncronas externas
- Escolher um bom limiar de particionamento
  - Emails
  - Blogs
  - Maps

Modelo de dados



Mapeando para BigTable

- Nome_Tabela + Nome_Propriedade = sem colisão
- Linha da tabela root armazena metadados do log

Row key	User. name	Photo. time	Photo. tag	Photo. _url
101	John			
101,500		12:30:01	Dinner, Paris	...
101,502		12:15:22	Betty, Paris	...
102	Mary			

Algoritmos – Leitura Corrente

1. verificar com o coordinator: entity group está atualizado localmente?
2. verificar a posição mais atualizada e consistente do log. Escolher a replica mais atualizada possível
3. *Catchup*
4. A replica então é validada pelo coordinator como estando atualizada.
5. A leitura é finalmente realizada. Em caso de falha o catchup é realizado novamente.

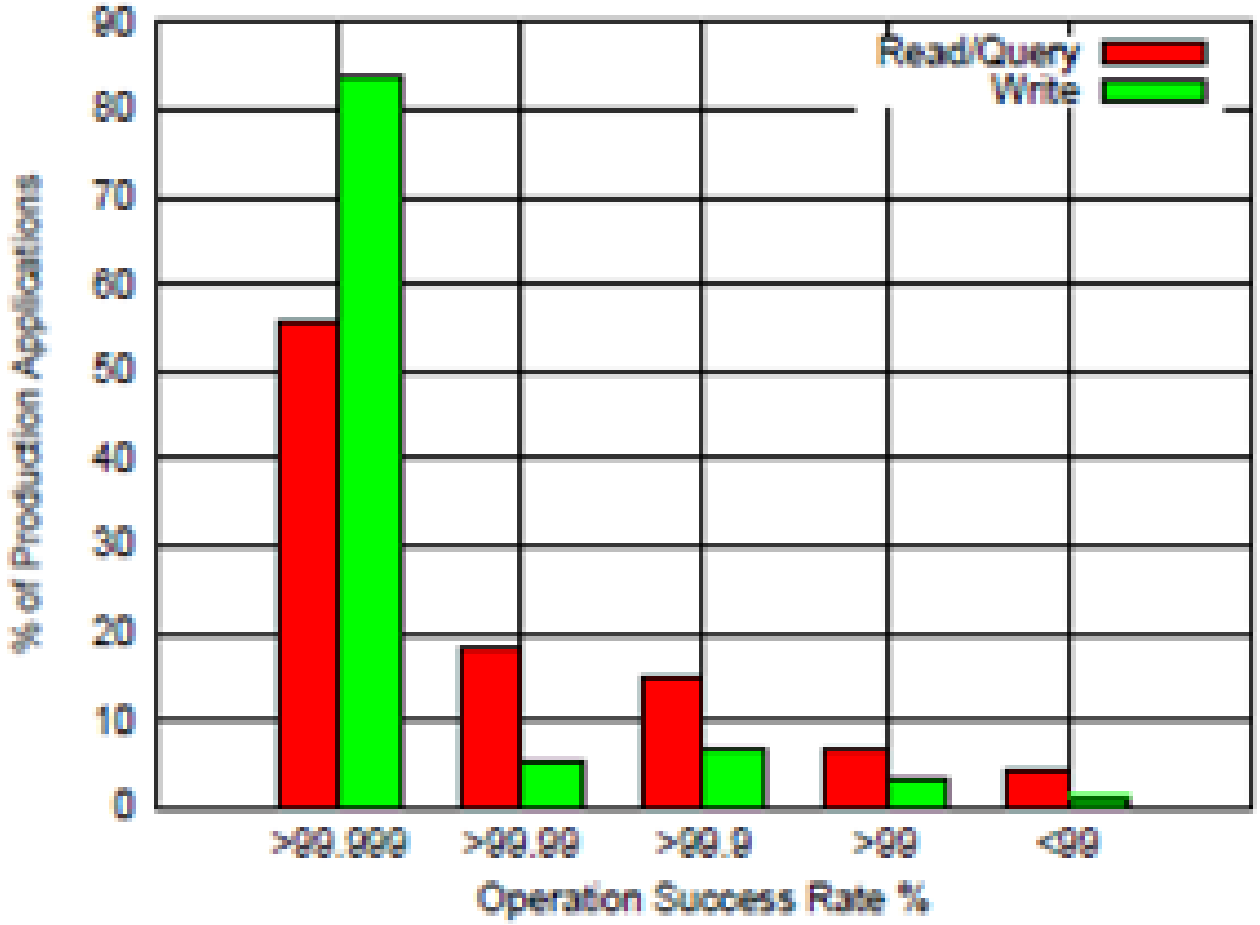
Algoritmos - Escrita

1. Aceitação do novo líder: Sim = Passo 3; Não = Passo 2;
2. Preparação do algoritmo de Paxos; Novo valor
3. Todas as replicas devem aceitar o novo valor. Se a maioria falhar: Passo 2;
4. Para as replicas que não aceitaram, o coordinator deve ser invalidado
5. Executar as alterações nos dados efetivamente

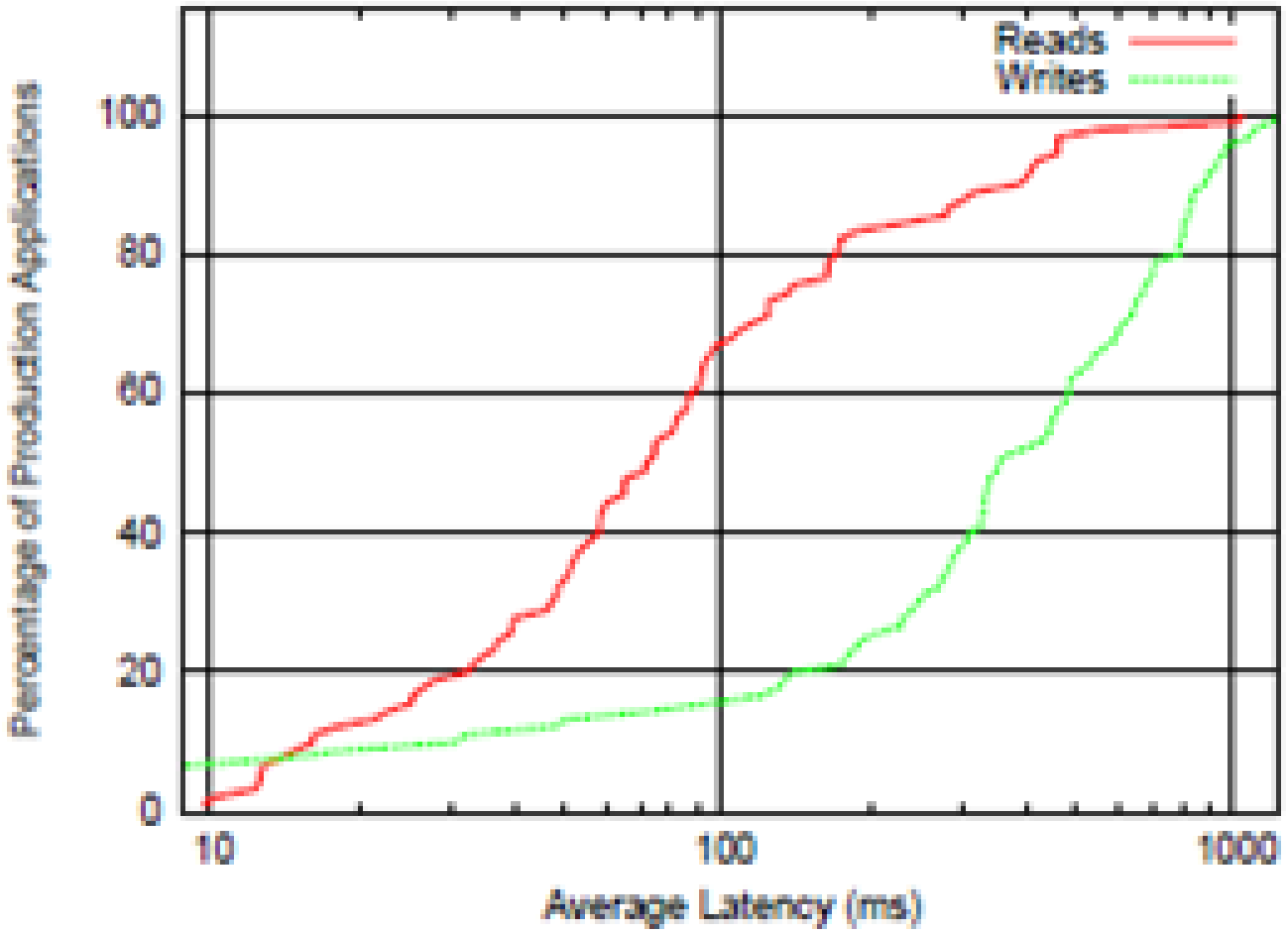
Tratamentos de falhas

- Falhas de replicas Full
  - Desviar o tráfego da replica
  - Desabilitar o coordinator
  - Desabilitar a replica 

Análises - Disponibilidade



Análises - Latência



Conclusão

- Solucionar o problema da crescente demanda da internet
- Aplicações críticas
- Tendência de combinação de abordagens

Referências

- [1] J. Baker, C. Bond, J. C. Corbett, JJ Furman, A. Khorlin, J. Larson, J-M. Léon, Y. Li, A. Lloyd, V. Yushprakh – Google Inc., "Megastore: Providing Scalable, Highly Available Storage for Interactive Services" in CIDR 2011, Asilomar, California, USA: Janeiro 9-12, 2011.
- [2] R. M. Toth, "Abordagem NoSQL – Uma real alternative" Sorocaba, São Paulo, Brasil: Abril 13, 2011.
- [3] P. Bernstein, "Google Megastore" in http://perspectives.mvdirona.com Julho 10, 2008.
- [4] J. Hamilton, "Google Megastore: The Data Engine behind GAE" in http://perspectives.mvdirona.com Janeiro 9, 2011.
- [5] T. Hoff, "Google Megastore - 3 Billion Writes And 20 Billion Read Transactions Daily" in http://highscalability.com Janeiro 11, 2011
- [6] K. Finley, "Google Announces High Replication Datastore for App Engine" in www.readwriteweb.com Janeiro 6, 2011
- [7] A. Girbal, "Two-phase commit" in www.mongodb.org Março 14, 2011
- [8] C. Lima, "Conceito de Surrogate Key – Chaves Substitutas" Março 9,2011
- [9] L. Lamport, "Paxos made simple" Novembro 1, 2011

Perguntas