



---

# Computação no *Alice* e grid

---

Alexandre Suaide  
IF-USP



# Resumo

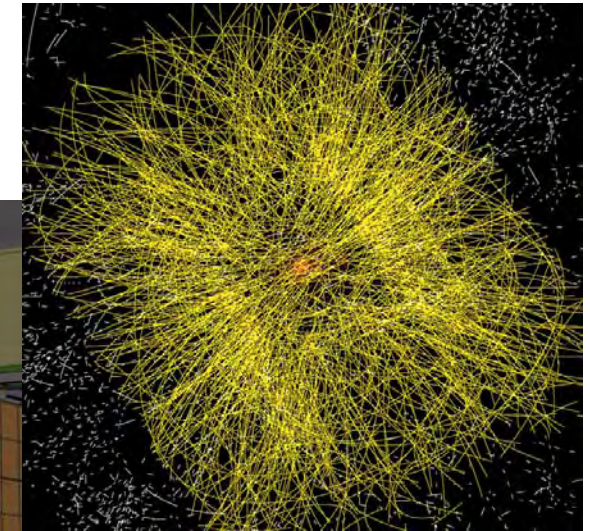
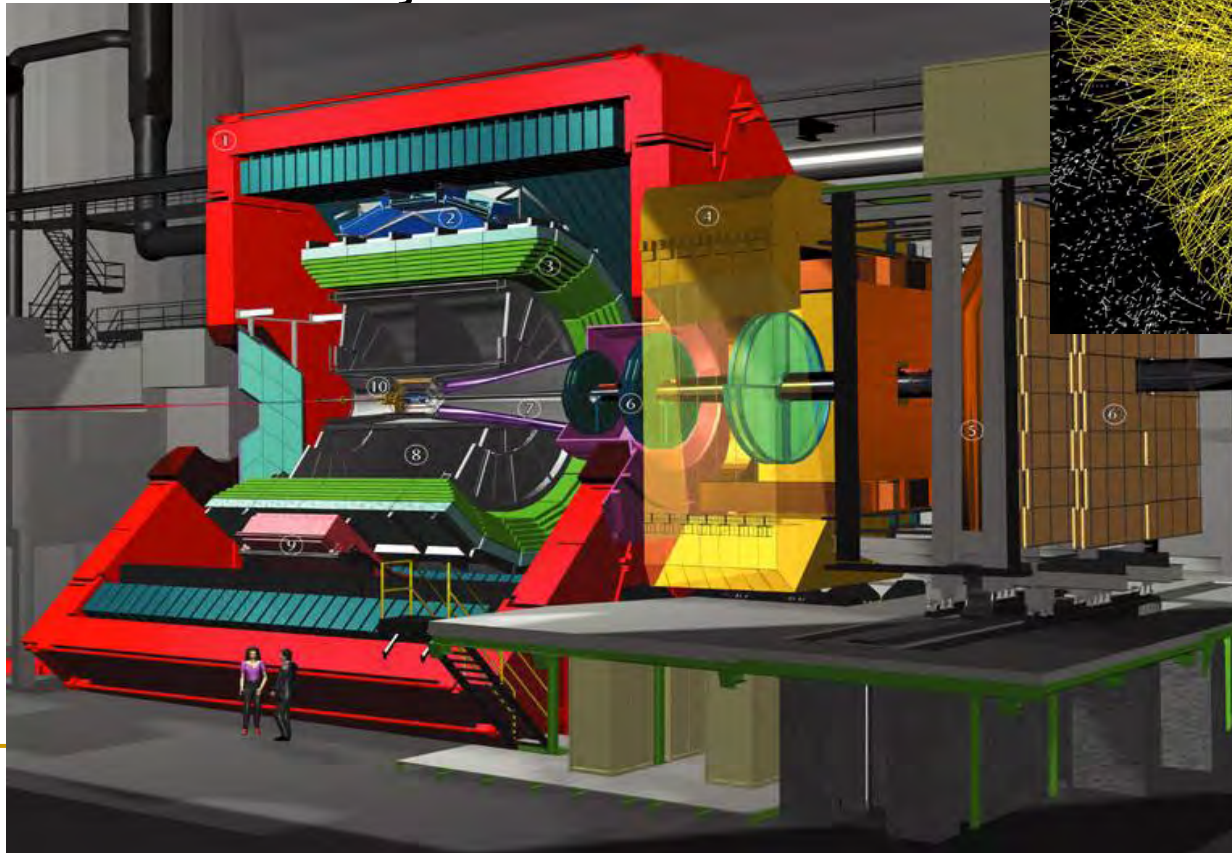
---

- Alguns números importantes
- Requerimentos (dados e simulações)
  - Processamento
  - Armazenamento
- Estrutura computacional
  - GRID
- Nossas necessidades



# Alice

- Um dos 4 experimentos no LHC
  - + 1000 colaboradores
  - + 100 instituições





# Parâmetros esperados durante operação

- Taxa de aquisição de eventos
  - p+p ~ 100 Hz (médio, pode ir a 500 Hz por curtos intervalos de tempo)
  - Pb+Pb ~ 100 Hz (em média, considerando ~12.5 MB/evento)
  
- Taxa para armazenamento (largura de banda)
  - p+p ~ 20 MB/s
  - Pb+Pb ~1.25 GB/s
  
- Tomada de dados
  - p+p ~ 10 meses/ano
  - Pb+Pb ~ 1 mês/ano
  
- Número de eventos adquiridos
  - p+p ~  $10^9$ /ano
  - Pb+Pb ~  $10^8$ /ano



# Requerimentos para processamento de dados

- **Reconstrução de eventos**
  - CPU/evento (médio)
    - p+p ~ 5.4 kSI2k x s
    - Pb+Pb ~ 675 kSI2k x s
  - Número de eventos
    - p + p ~  $10^9$
    - Pb + Pb ~  $10^8$
  - CPU/ano (considerando 3 passos durante a produção)
    - p + p ~ 800 kSI2k
    - Pb + Pb ~ 7800 kSI2k
  
- CPU TOTAL para eventos reais ~ 8.6 MSI2k

**1 kSI2k equivale ao Benchmark (aproximado) de um PIV a 3.0 GHz**



# Requerimentos para análise de dados

- **Análise de dados**
  - CPU/evento
    - p + p ~0.3 kSI2k x s
    - Pb + Pb ~34 kSI2k x s
  - Número de eventos
    - p + p ~  $10^9$
    - Pb + Pb ~  $10^8$
  - CPU/ano (15 PWG, cada um analisando 1% dos dados por dia)
    - p + p ~ 800 kSI2k
    - Pb + Pb ~ 10000 kSI2k
  
- **CPU TOTAL para análise de dados ~ 10.5 MSI2k**



# Requerimentos para simulações

- Simulação dos eventos
  - CPU/evento
    - $p + p \sim 27 \text{ kSI2k} \times s$
    - $Pb + Pb \sim 10125 \text{ kSI2k} \times s$
  - Número de eventos
    - $p + p \sim 10^8$
    - $Pb + Pb \sim 2 \times 10^5$
  - CPU/ano
    - $p + p \sim 80 \text{ kSI2k}$
    - $Pb + Pb \sim 70 \text{ kSI2k}$
- TOTAL CPU para simulações  $\sim 0.15 \text{ MSI2k}$



# Requerimentos para simulações

- Reconstrução dos eventos simulados
  - CPU/evento (médio)
    - p + p ~ 5.4kSI2k x s
    - Pb + Pb ~675 kSI2k x s
  - CPU/ano (assumindo 2 passos na produção)
    - p + p ~ 40 kSI2k
    - Pb + Pb ~2530 kSI2k
  - TOTAL CPU para reconstrução (1 passo) ~ 2.6 MSI2k
  
- TOTAL CPU para dados simulados ~ 2.7 MSI2k



# Armazenamento de dados

---

- Dados brutos
  - Tamanho de evento
    - $p + p \sim 0.2 \text{ MB}$
    - $Pb + Pb \sim 12.5 \text{ MB (médio)}$
  - Número de eventos
    - $p + p \sim 10^9$
    - $Pb + Pb \sim 10^8$
  - Volume de dados
    - $p + p \sim 0.2 \text{ PB}$
    - $Pb + Pb \sim 1.3 \text{ PB}$
  
- TOTAL  $\sim 3 \text{ PB}$  (incluindo fator de replicação = 2)



# Armazenamento de dados

## ■ Produções

### □ Tamanho de evento

- p + p ~ 0.02 MB
- Pb + Pb ~ 1.25 MB (medio)

### □ Número de eventos

- p + p ~  $10^9$
- Pb + Pb ~  $10^8$

### □ Número de passos da produção =

### □ Volume de dados (ESD + AOD + TAG)

- p + p ~ 0.25 PB
- Pb + Pb ~ 1.4 PB

## ■ TOTAL ~ 4.1 PB (incluindo AOD) e 3 (ESD))

Event Summary Data  
O evento reconstruído  
Tracks, V0's, kinks, etc

Informações globais sobre os eventos.

Analysis Object Data  
Informações filtradas para cada PWG.

o = 2.5



# Armazenamento de simulações

- Dados brutos
  - Tamanho de evento
    - $p + p \sim 0.4$  MB
    - $Pb + Pb \sim 300$  MB (médio)
  - Número de eventos
    - $p + p \sim 10^8$
    - $Pb + Pb \sim 2 \times 10^5$
  - Volume de dados
    - $p + p \sim 0.04$  PB
    - $Pb + Pb \sim 0.06$  PB
  
- TOTAL  $\sim 0.1$  PB



# Armazenamento de simulações

- Simulações brutas
  - Tamanho de evento
    - p + p ~ 0.04 MB
    - Pb + Pb ~ 2.5 MB (médio)
  - Número de eventos
    - p + p ~  $10^8$
    - Pb + Pb ~  $2 \times 10^5$
  - Volume de dados (incluindo replicação de 2.5)
    - p + p ~ 0.01 PB
    - Pb + Pb ~ 0.6 PB
  
- TOTAL ~ 0.7 PB



# Resumindo...

## ■ Processamento

■	Reconstrução de eventos reais	8.6 MSI2k
■	Análise de dados	10.5 MSI2k
■	Simulação (total)	2.7 MSI2k
□	<b>Total de processamento</b>	<b>21.8 MSI2k</b>

## ■ Armazenamento

■	Dados brutos	3.0 PB
■	Produções	1.6 PB
■	Simulações	0.8 PB
□	<b>Total em fita</b>	<b>5.0 PB</b>
□	<b>Total em disco</b>	<b>6.5 PB</b>



# Mais requerimentos: tempo

- Processamento de dados (p + p)
  - Calibração e alinhamento – (quase) online
  - Primeiro passo de reconstrução durante aquisição
    - Medida de propriedades globais rápido
    - Ajuste fino da reconstrução
  - Segundo passo da produção logo em seguida
- Processamento de dados (Pb + Pb)
  - Calibração e alinhamento durante aquisição
  - Primeiro passo de reconstrução ~ 4 meses
  - Segundo passo de reconstrução ~ 6 meses

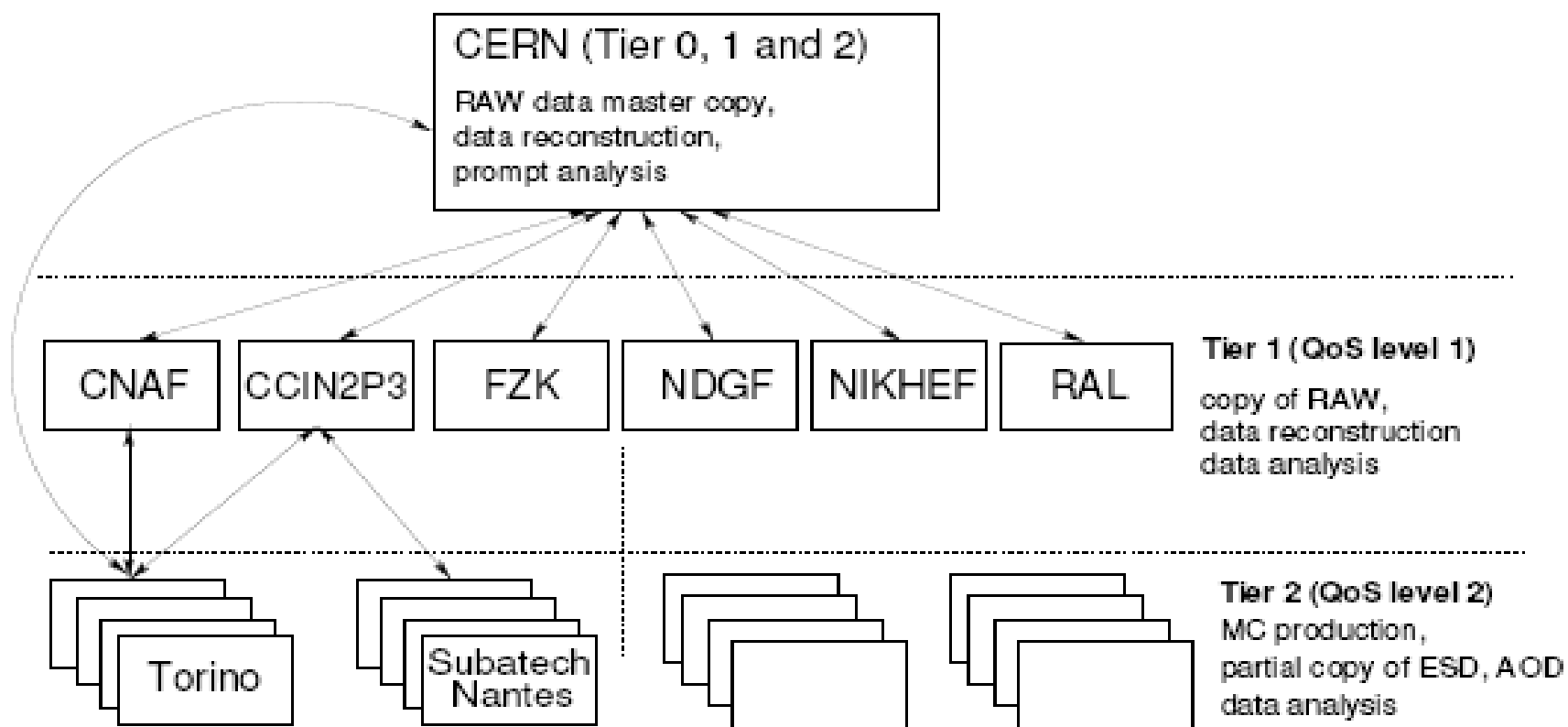


# Como suprir essas necessidades?

- Solução: computação distribuída (GRID)
  - Usar computadores disponíveis em centros em todo mundo, em uma estrutura de árvore (tier)
  - Produção, simulação e análise de dados (interativa ou batch) pode usar todos esses recursos através de ferramentas comuns de grid
    - Por enquanto, utilizar o sistema caseiro
      - AliEn – Alice Environment



# Estrutura hierárquica





# AliEn – Alice Environment

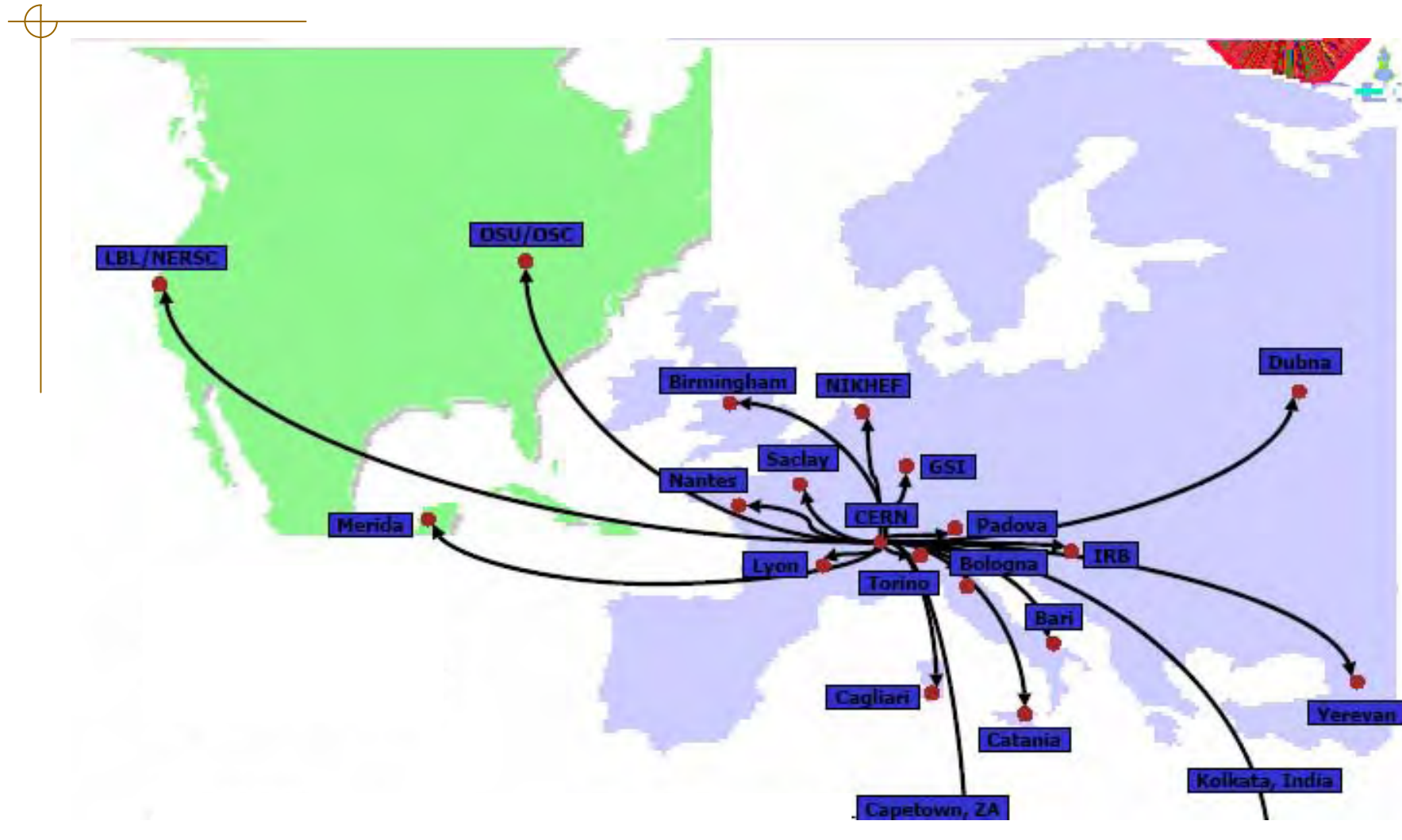
<http://alien.cern.ch>



- Sistema de ferramentas de grid desenvolvidas pelo Alice
  - Em uso há mais de 4 anos
  - Baseado em serviços web com protocolos padrão
  - Baseado em software livre disponível
    - SASL/OpenSSL/OpenCA – autenticação
    - Globus/GSS – autenticação em grid
    - Condor – jobs
    - OpenLDAP – configuração
    - Apache, MySQL, Bbftp, etc
    - Menos de 5% é puramente nativo do AliEn
  - + 500.000 jobs executados utilizando esse sistema



# AliEn@GRID



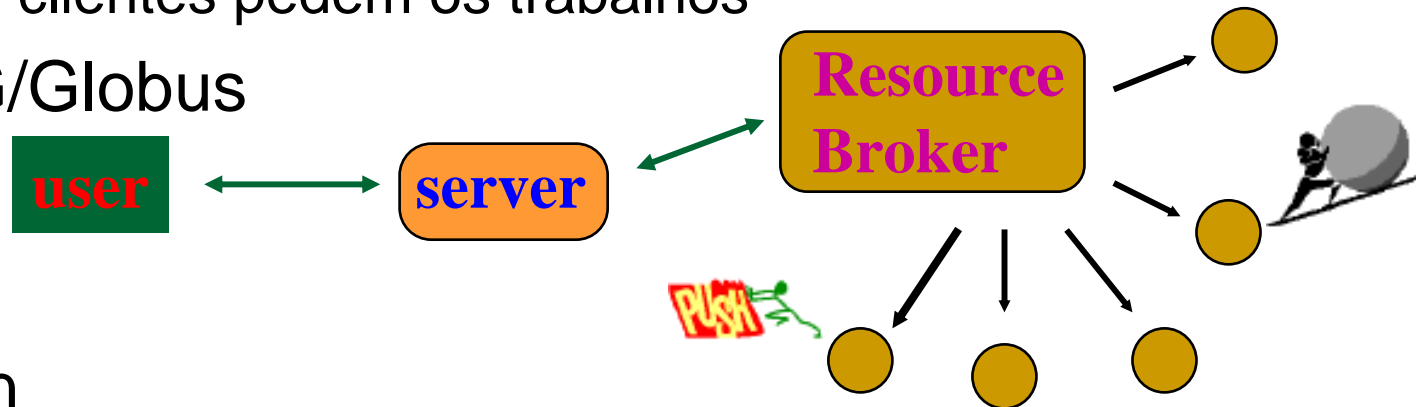


# AliEn protocolo básico

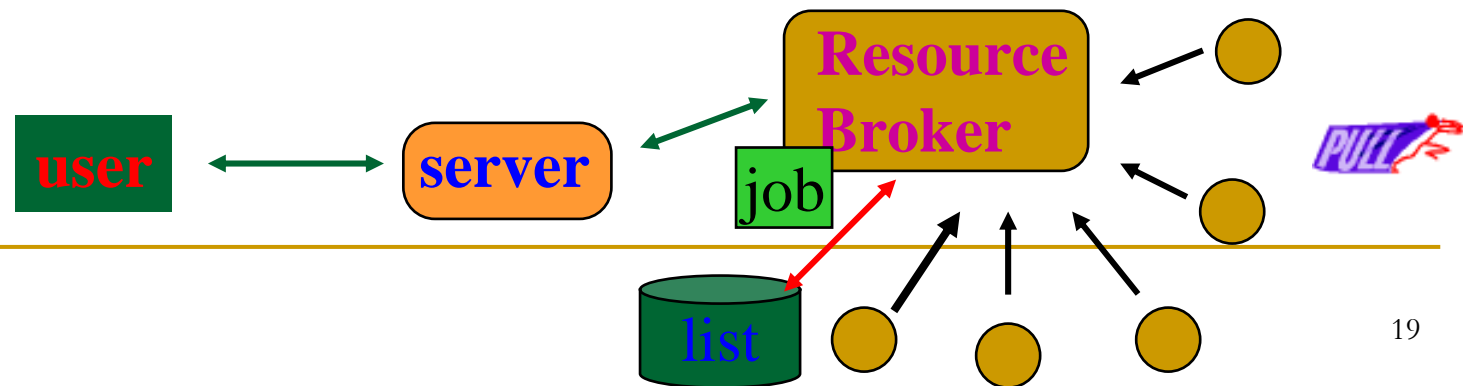
- A maior diferença entre Alien e outros protocolos de grid é o uso de protocolos do tipo 'pull' ao invés de 'push'

- Os clientes pedem os trabalhos

- EDG/Globus



- AliEn





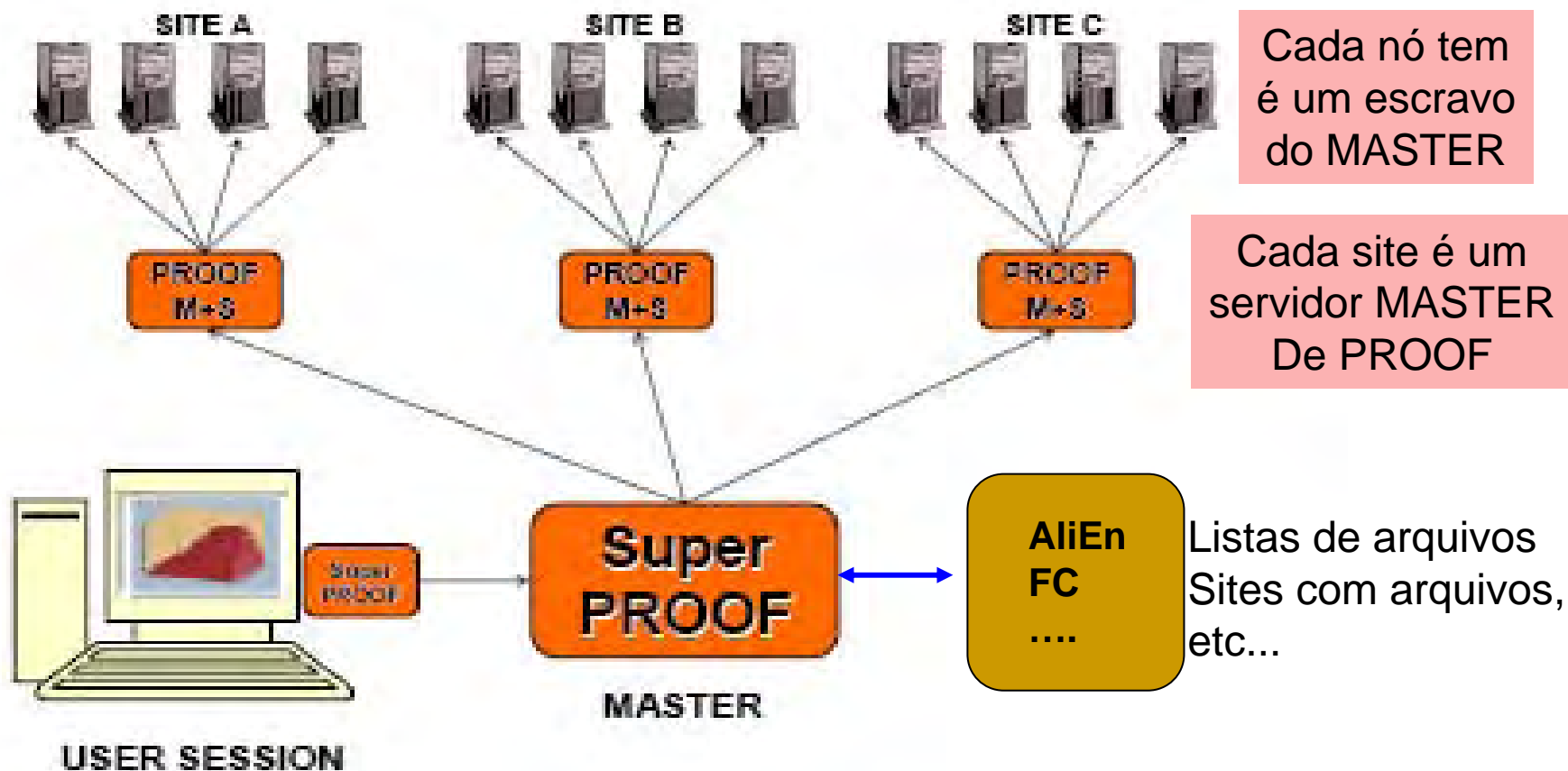
# Análise de dados no Alice

---

- Análise de dados é uma atividade caótica
    - Em todos sentidos (tempo, algoritmos, acesso a dados, etc)
  - Necessita amplo instrumental para análise
    - ROOT
  - Será realizada, na maioria, nos sites Tier-1 e Tier-2
    - Deve fazer uso eficiente de ferramentas de GRID
    - Acoplar AliEn + ROOT (PROOF)
-



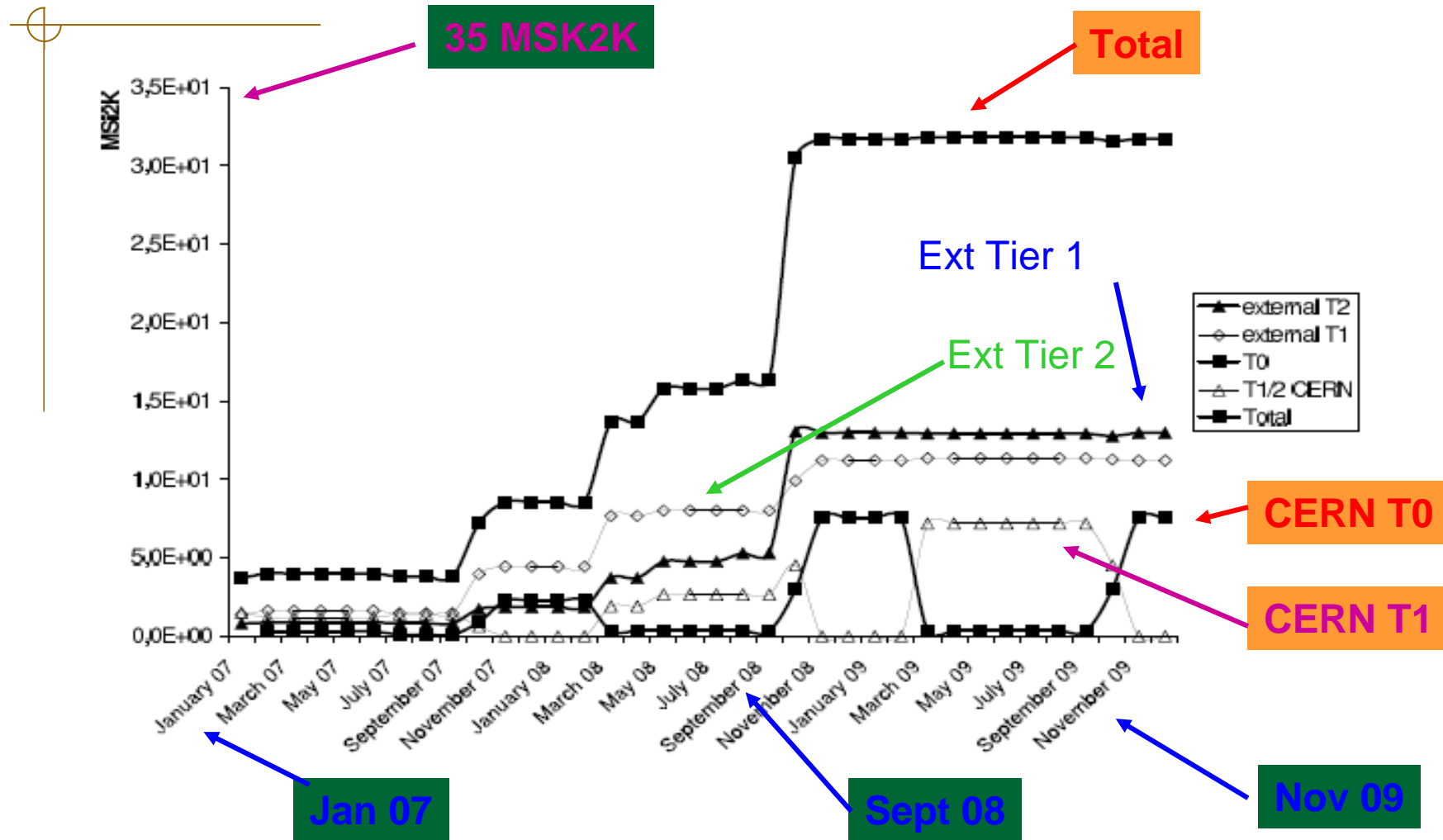
# PROOF no Alice



Uso de protocolos do tipo 'pull', ou seja, os diferentes nós requisitam trabalho do servidor MASTER



# Distribuição de CPUs





# TIER-1

---

- Hipóteses:
  - Contribui com reconstrução, simulação e análise
  - Aproximadamente com 40% de todo poder de processamento
  - Número de Tier-1s = 3 (France, Germany, Italy),
  
- Poder de processamento (médio) em cada Tier-1 **3 MSI2K**
  
- Capacidade de disco de cada Tier-1 **1 PB**
  
- Capacidade de fita em cada Tier-1 **1 PB**
  
- Largura de banda de rede para o CERN **140 Mb/s**



# TIER-2

---

## ■ Hipóteses:

- Contribui com simulações e análise de dados
- Aproximadamente 33% de toda capacidade de processamento
- Número de Tier-2 = 20

- Poder de processamento de cada Tier-2 **0.36 MSI2K**
- Capacidade de disco de cada Tier-2 **0.15 PB**
- Largura de banda de rede para Tier-1 **100 Mb/s**



# O cluster local em São Paulo

- O nosso grupo vem atuando intensivamente na implementação do grid para o experimento STAR, em BNL.
- O nosso cluster está totalmente implementado no OSG e executa, constantemente, para o STAR:
  - Reconstrução de eventos
  - Simulações
  - Análise de dados em desenvolvimento
- Porém, isso é feito em circunstâncias de testes
  - Cluster com pouco poder de processamento para fazer diferença no STAR
  - É mais uma prova de princípio



# Cluster local

---

- **Processamento**
    - Pouco poder de processamento (~ 0.02 MSI2K)
      - Compartilhado com necessidades locais
  - **Armazenamento**
    - Disco distribuído. Acesso central somente para /home
    - 2 TB de capacidade total
      - Usada basicamente para dados do STAR
  - **Sistema de jobs**
    - Scheduler do STAR + SGE
      - Os jobs são distribuídos dependendo da disponibilidade de CPU e onde os dados estão armazenados
-



# Conectividade

- O maior fator limitante para executar operações via grid, principalmente reconstrução e simulação é a largura de banda disponível
  - Conexão feita através da rede interna do IF-USP + CCE
    - Compartilhamento com todo IFUSP
    - Rede instável
    - Pouca largura disponível
      - Transferências volumosas praticamente impossíveis



# Atualização do cluster

---

- **Processamento e armazenamento**
  - Pedido FAPESP
    - Cluster 0.10-0.15 MSI2k (80-150 CPUS)
    - Armazenamento 0.10 – 0.15 PB
    - Projeto recentemente aprovado
      - Ainda não sabemos o montante aprovado
- **Rede**
  - Melhorar desempenho da rede do IF-USP
    - Porém, continuaremos compartilhando banda
  - Usar conexão dedicada
    - Emprestar uma das fibras ópticas disponíveis no SPRACE
      - Otimiza conexão entre os dois clusters



# Sumário

---

- O armazenamento, reconstrução e análise de dados do Alice é desafiador
    - Requer computação distribuída
  - O nosso grupo possui experiência em processamento e transferência de dados usando tecnologia de grid
  - Muito em breve o cluster local será atualizado e terá capacidade para processar e armazenar dados do Alice
-