MC: Workloads Ciêntificos na Computação em Nuvem

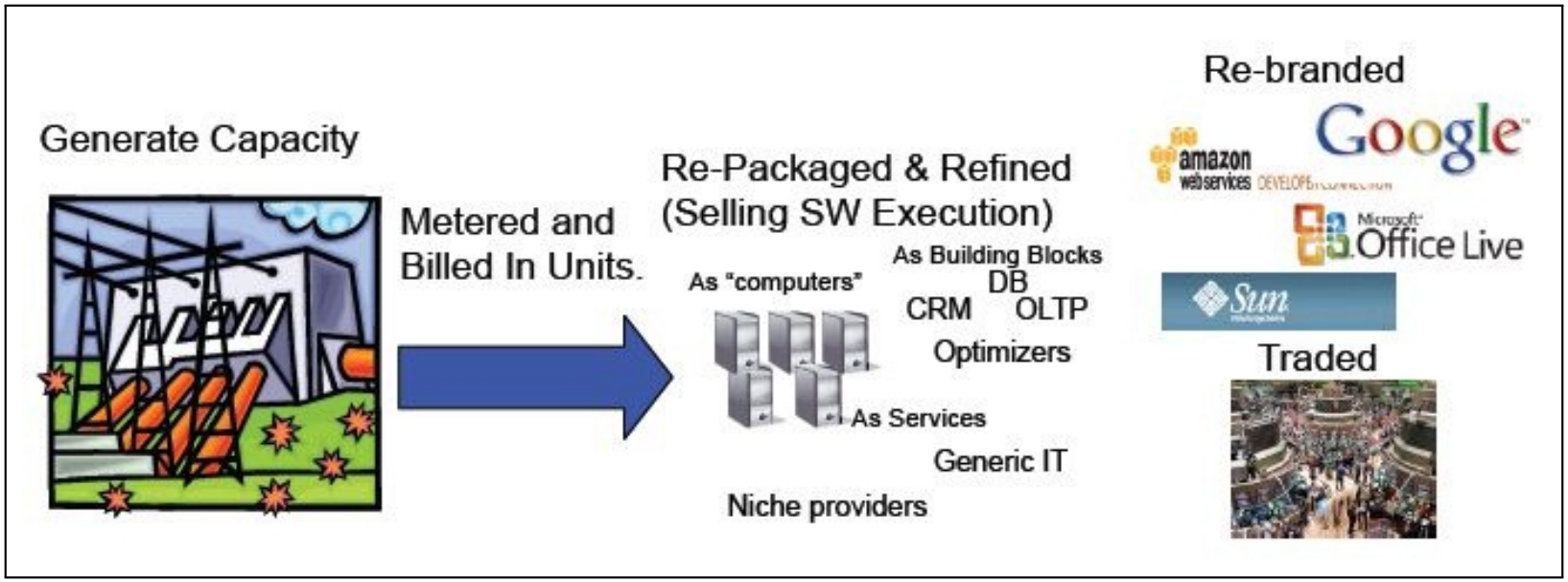
Uma Introdução

Cesar Marcondes e Hermes Senger

Universidade Federal de São Carlos

ERAD-SP - 29 Julho 2011

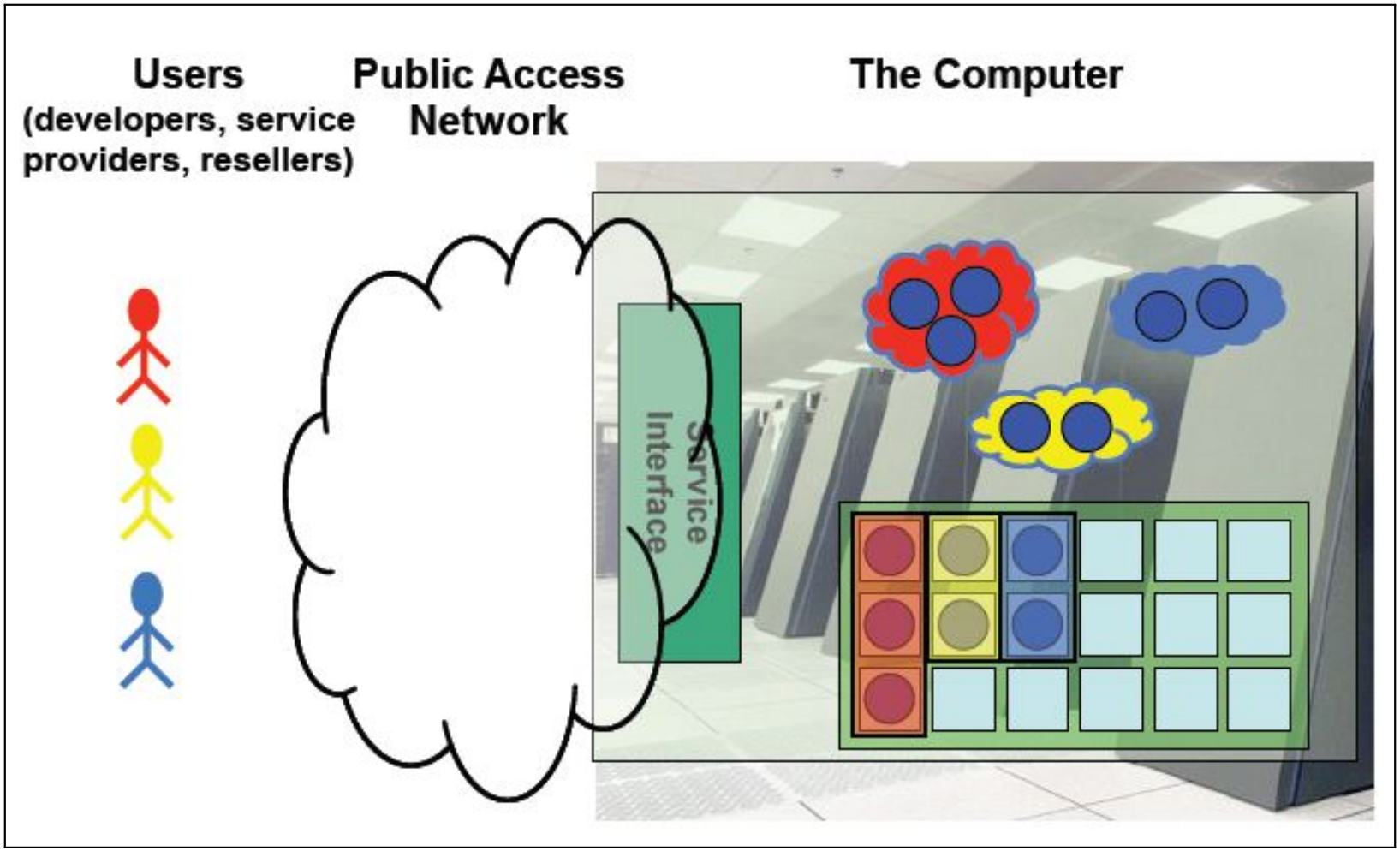
Computação Utilitária



- Produzida através de Capacidade Instalada
- Distribuida
- Empacotada em Software

Resultado: Unidades Genéricas de Commodities Computacionais que são consumidas e utilizadas baseado em um valor provisionado

Uma Visão Abstrata do Sistema

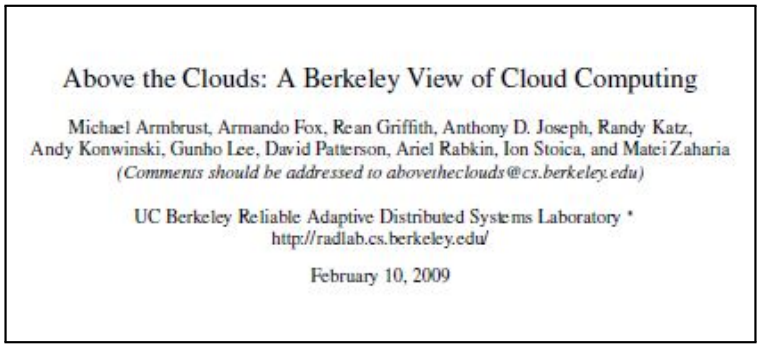


Uma Definição - Computação em Nuvem

.. a broad array of
web-based services aimed at
allowing users to obtain a
wide range of functional capabilities
on a pay-as-you-go basis
that previously required tremendous
hardware/software investments
and professional skills to acquire.

Irving Wladawsky-Berger
Chairman Emeritus, IBM Academy of Technology

Computação em Nuvem - Não é Novo mas Gera Hype



Above the Clouds: A Berkeley View of Cloud Computing

Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz,
Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia
(Comments should be addressed to abovetheclouds@cs.berkeley.edu)

UC Berkeley Reliable Adaptive Distributed Systems Laboratory *
http://radlab.cs.berkeley.edu/

February 10, 2009

- Ilusão de recursos computacionais INFINITOS e SOB DEMANDA
- Nenhum COMPROMISSO antecipado por parte dos usuários
- PAGAMENTO pelo USO de recursos a CURTO prazo, conforme necessário

Modelos de Computação em Nuvem - IAAS

InfraStructure As A Service

Consiste no fornecimento de infraestrutura computacional (normalmente um ambiente de virtualização de plataforma) como um serviço, alavancando importante tecnologia e investimentos em data centers para fornecer TI como um serviço aos clientes.

Exemplos de InfraStructure As A Service

Modelos de Computação em Nuvem - PAAS

Platform As A Service

É o fornecimento de uma plataforma na nuvem, sobre a qual os aplicativos podem ser desenvolvidos e executados utilizando linguagens de programação e ferramentas suportadas pelo provedor do serviço.

Exemplos de Platform As A Service

Modelos de Computação em Nuvem - SAAS

Software As A Service

É um modelo de distribuição de software no qual as aplicações são hospedadas por um fornecedor ou prestador de serviços e disponibilizadas aos clientes através de uma rede, geralmente a Internet.

Modelos de Computação em Nuvem - SAAS

Exemplos de Software As A Service

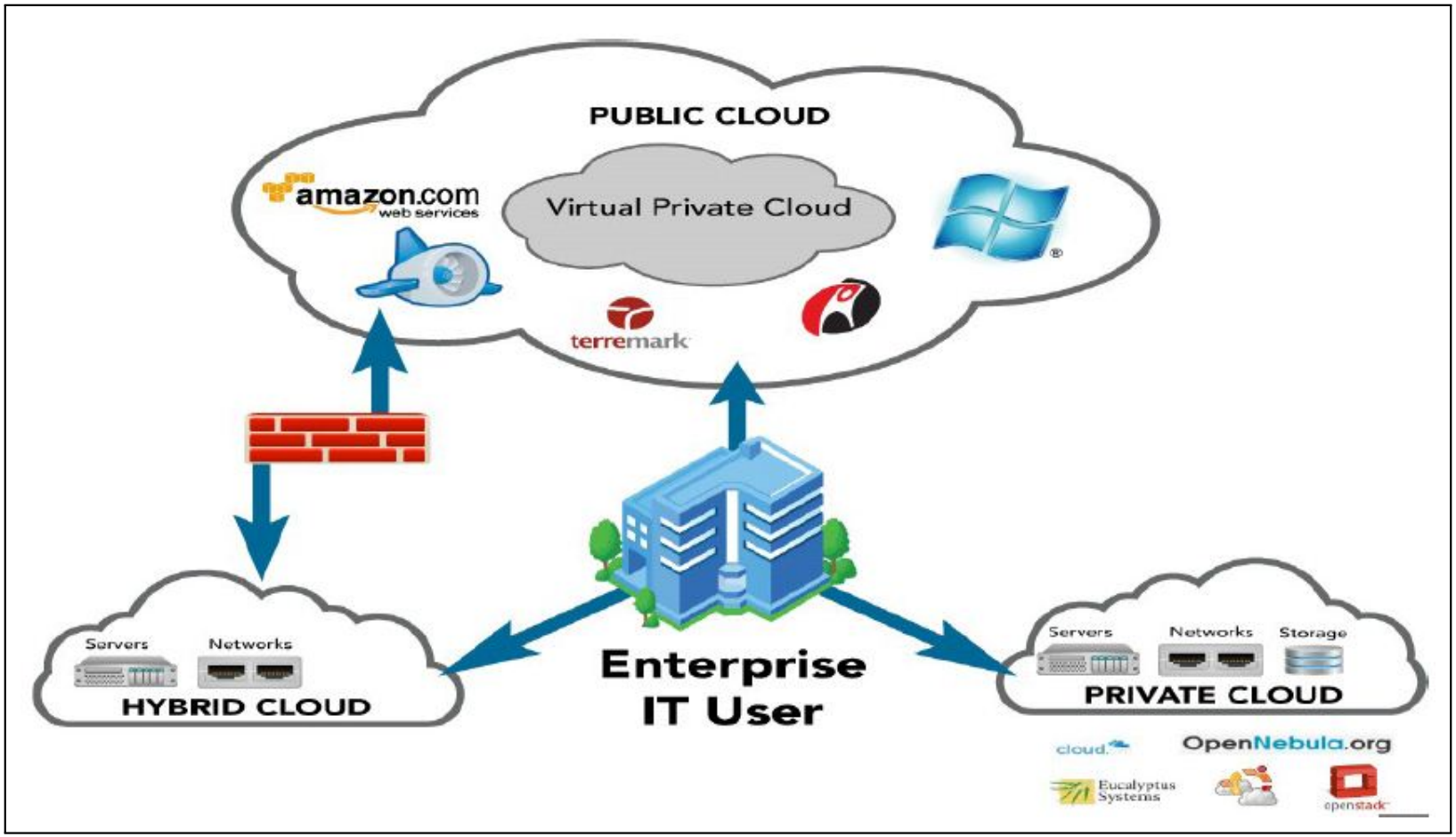
Modelos de Entrega

- Nuvem Pública: baseia-se no modelo padrão de Computação em Nuvem, no qual recursos de um prestador de serviço, tais como aplicações e armazenamento, estão disponíveis ao público em geral ou a um grande grupo industrial e são controlados por uma organização que vende os serviços de nuvem através da Internet.

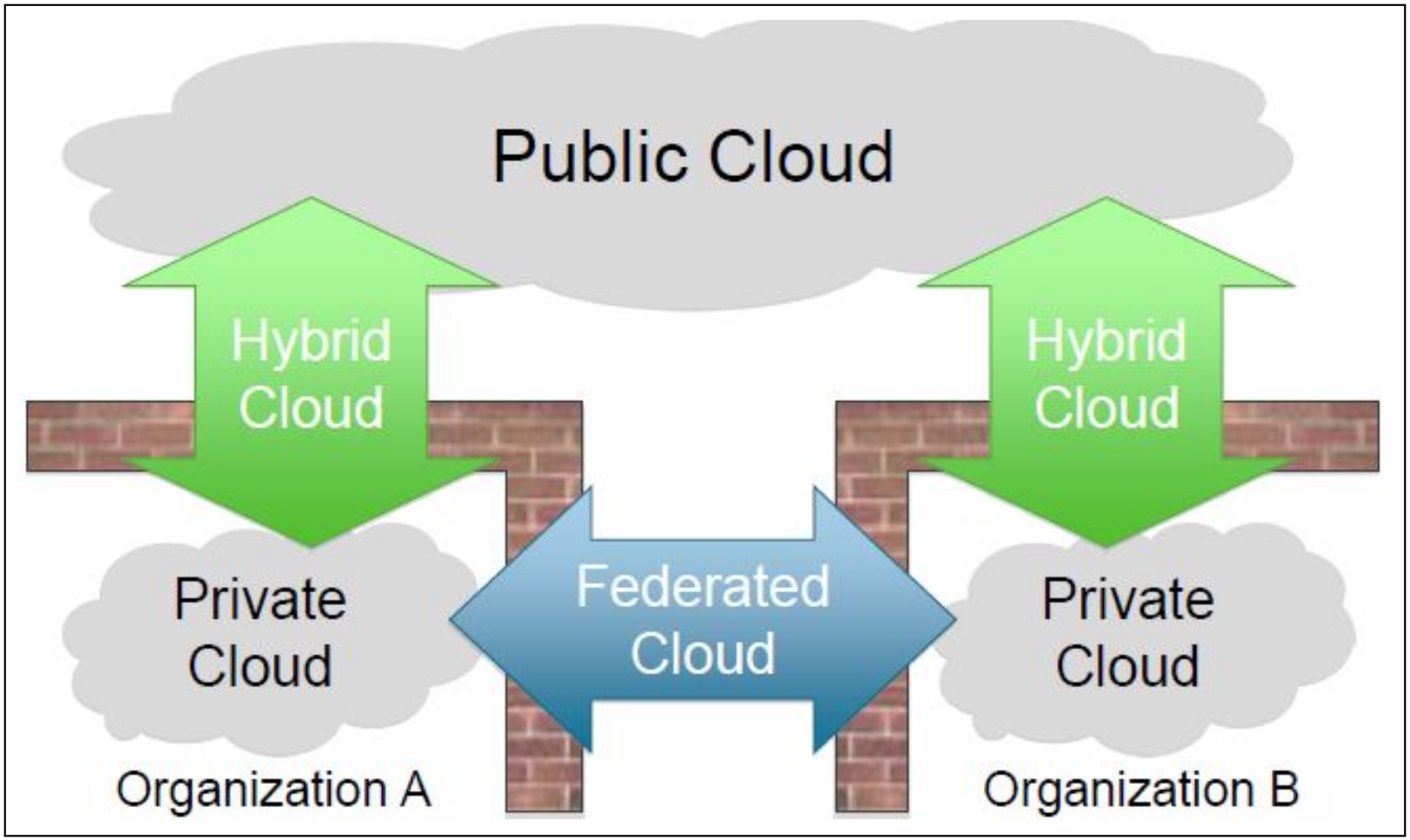
- Nuvem Privada: é um termo de marketing para uma arquitetura de computação de propriedade que fornece serviços hospedados para um número limitado de pessoas por dentro de um firewall.

- Nuvem Híbrida: uso de nuvens públicas e privadas em conjunto. Organizações que utilizam esse modelo podem hospedar aplicações críticas em nuvens privadas e aplicações com, relativamente, menos preocupações de segurança na nuvem pública.

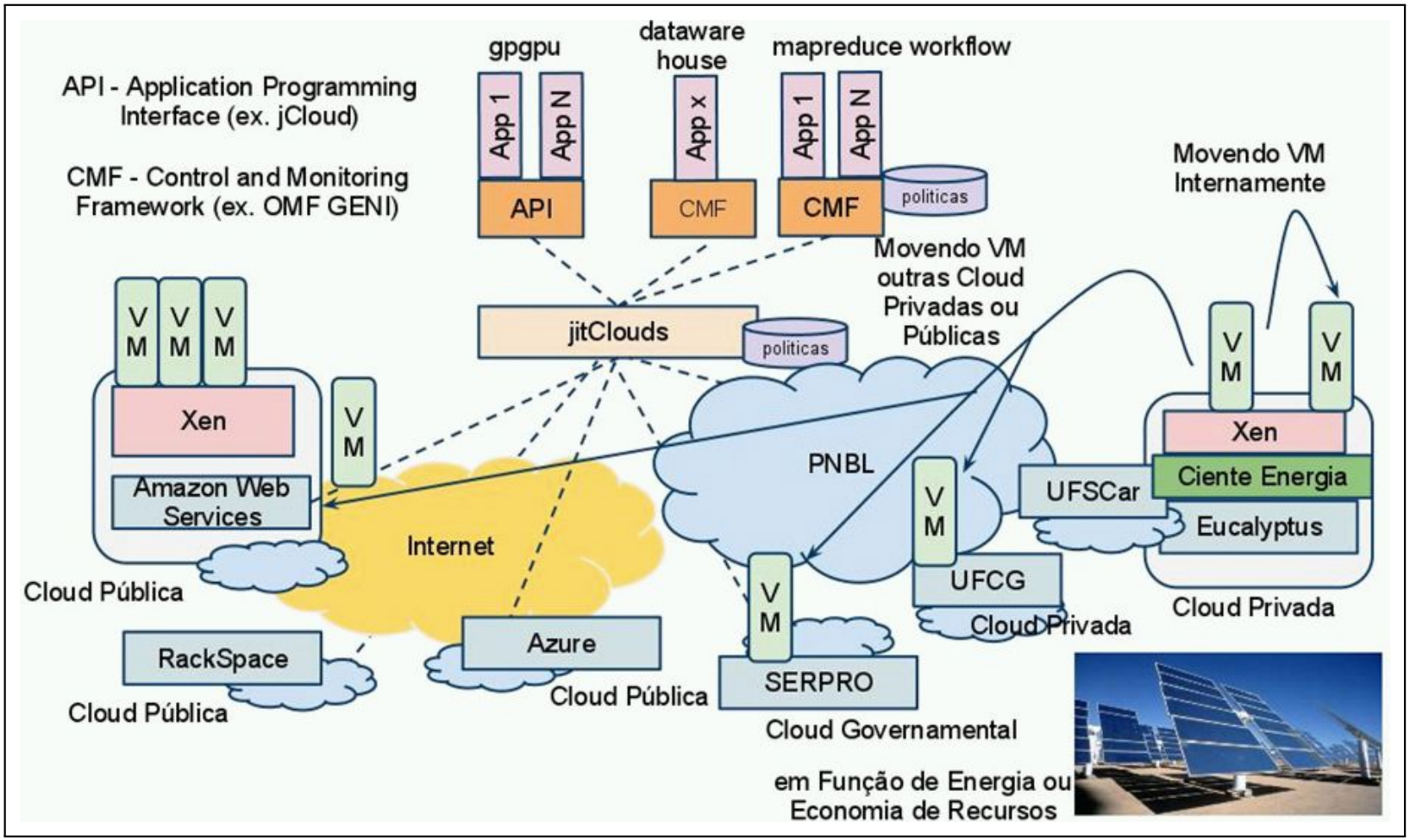
Modelos de Entrega



Modelos de Entrega - Federated



jitClouds - Just in Time Cloud Federation (RNP/MCT)



Open Source Cloud - Virtualização

Open Source

- Xen, Xen Cloud Platform (XCP)
- KVM – Kernel-based Virtualization
- VirtualBox - Oracle supported Virtualization Solutions
- OpenVZ - Container-based, Similar to Solaris Containers or Zones
- LXC – Userspace chrooted installs

Proprietary

- VMware
- Citrix Xenserver
- Microsoft Hyper-V
- Oracle VM

OpenVZ

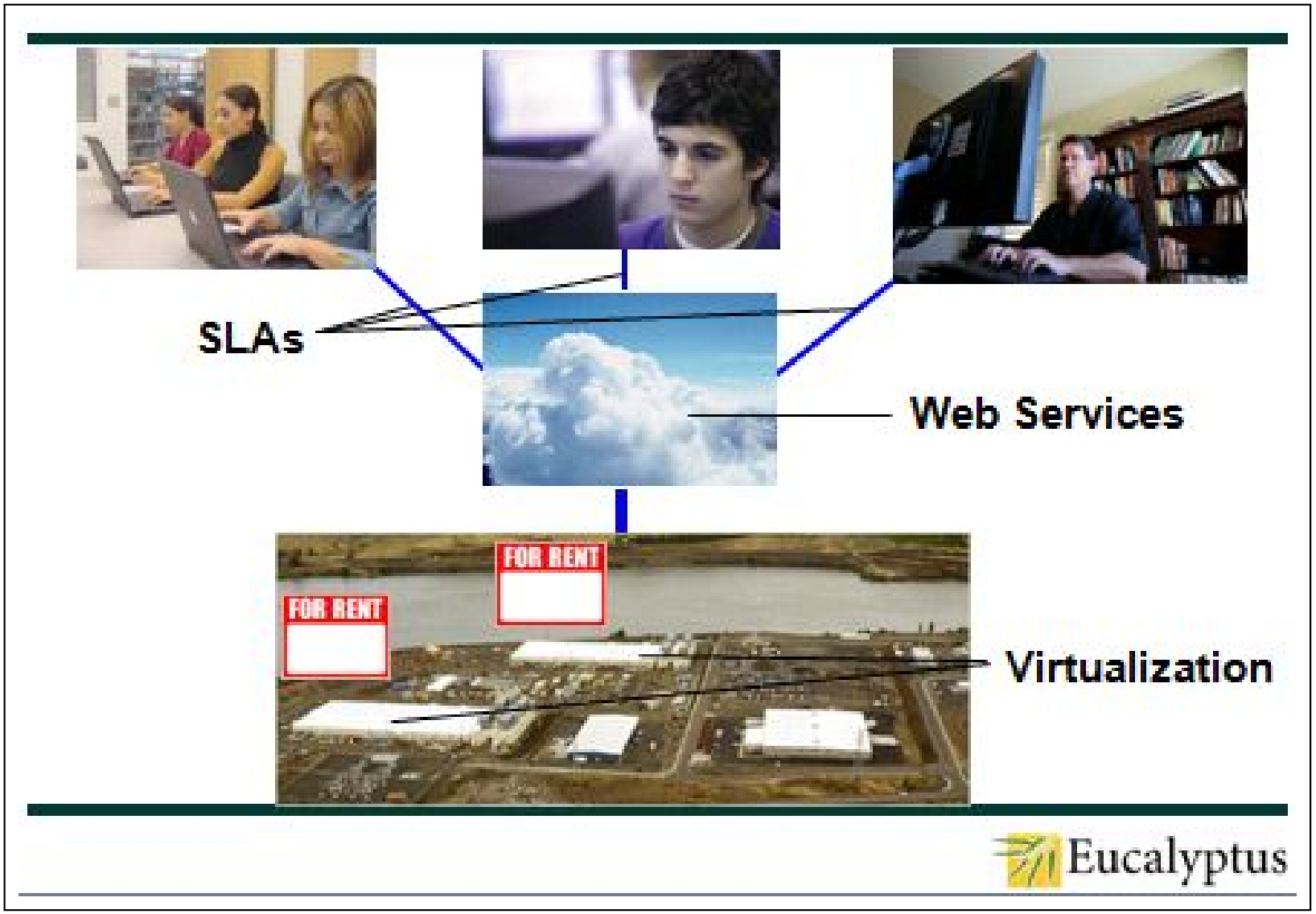
Open Source Cloud - Infrastructure As A Service

	Year Started	License	Hypervisors Supported
cloud.com	2010 (Development Since 2008)	GPLv3+	Xenserver, XCP, VMware, KVM
Eucalyptus Systems	2008	GPLv3	Xen, KVM, VMware
openstack	2010 (Developed by NASA by Anso Labs previously)	Apache	Xen, KVM, Hyper-V
Ubuntu Enterprise Cloud	2009 (Karmic Koala)	GPLv3	Xen,KVM
abiquo	2009 (Development 2006)	LGPLv3	VMware ESX and ESXi, Microsoft Hyper-V, Xen, KVM and Virtual Box

Open Source Cloud - Platform As A Service

	Year Started	Sponsors	Platforms Supported
CLOUD FOUNDRY	2011	VMware	Spring, Rails, Sinatra, Node.js
OPENSHIFT PaaS by Red Hat Cloud	2011	Red Hat	Express – Ruby, PHP Python Flex – JBoss, Java EE6
WSO2 Stratos	2009	WS02	Tomcat, JBoss, Java EE6,

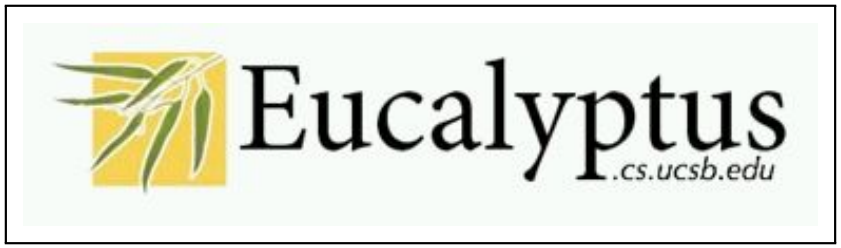
Eucalyptus


SLAs
Web Services
FOR RENT
FOR RENT
Virtualization
Eucalyptus

Caminho Inverso Cloud Pública para Cloud Privada

- Trata-se de uma infra-estrutura de larga escala disponível para aluguel
  - Sistema Operacional virtualizado (ex. Xen) provê isolamento da CPU
  - Provisionamento da Rede 'pelo próprio usuário' provê isolamento de rede
  - Abstrações sobre o Armazenamento Específico
- Serviço Completamente Self-Service
  - Acordos de Nível de Serviço são Anunciados (SLAs)
  - Requisições são aceitas e recursos são obtidos através de web services
  - Acesso aos recursos pelos Consumidos pode ser feito remotamente pela Internet
- Contabilização é baseada em Comércio Eletrônico
  - Transações baseadas na Web
  - Modelos de Negócio 'Pay-as-you-go' ou assinatura flat
  - SAC, reembolsos, problemas, etc.

Eucalyptus

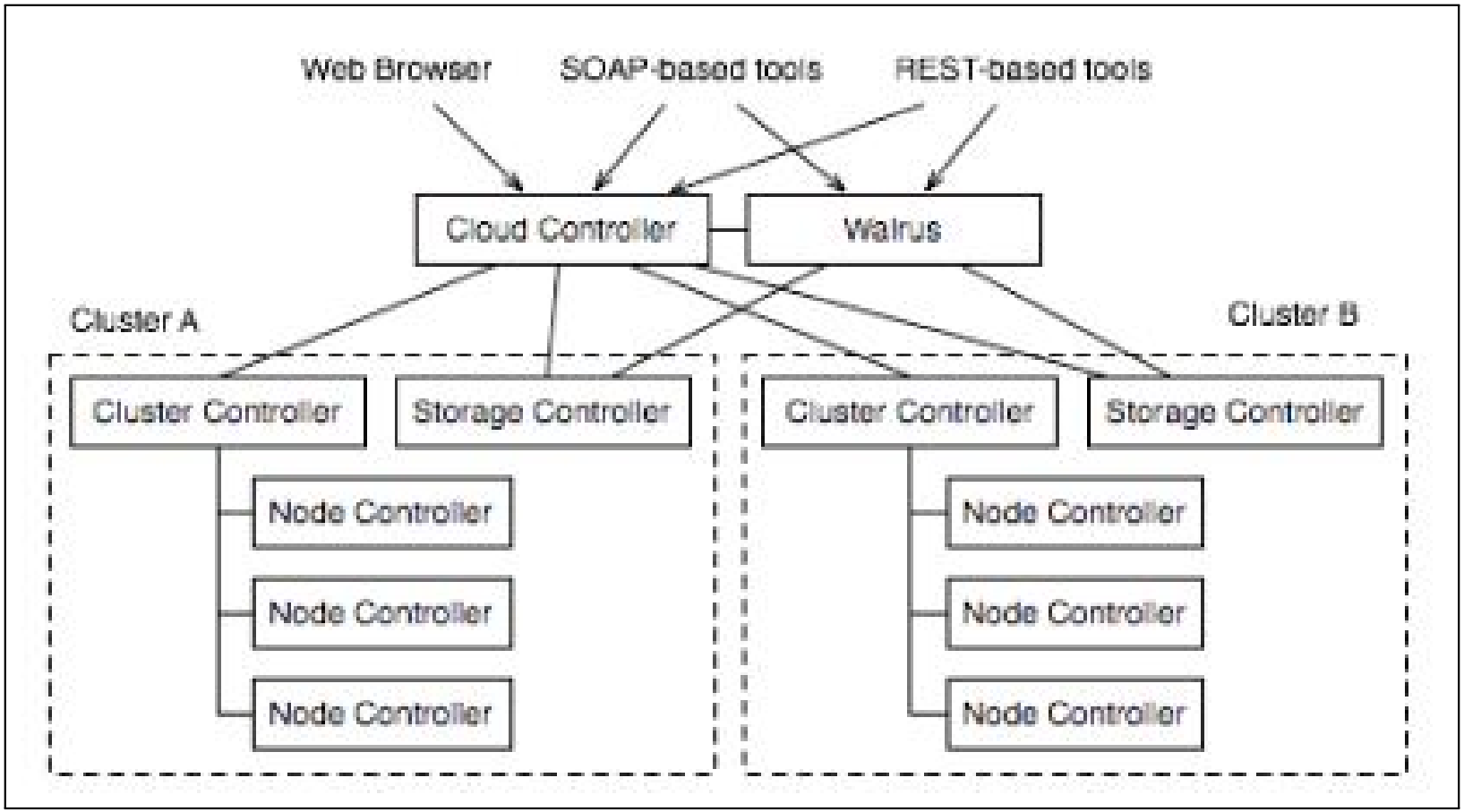


- Elastic Utility Computing Architecture Linking Your Programs To Useful Systems
  - Implementação baseada em Web services de uma infraestrutura de computação em nuvem (elastic/utility/cloud)
  - Armazenamento de Imagens Linux ala Amazon AWS
  - Como saber se isso é computação em nuvem? Emulando uma IaaS existente: Amazon AWS
  - Funciona como um camada de software (em alguns componentes) por cima da infraestrutura: De modo que a instalação não seja violada
  - Foco em Administração e Manutenção: 'Por que os Administradores de Sistema também são gente'

Coisas Boas com o Advento do Eucalyptus

- **Promover maior entendimento sobre computação em nuvem**: Provendo um veículo experimental que pode extender a nossa abordagem sobre o modelo de computação utilitária
- **Provendo Experimentação antes de adquirir serviços de nuvem comerciais**: Provê um desenvolvimento, debugging, e 'provas de conceito' de plataformas antes destas serem enviadas para Clouds Públicas
- **Homogeneizar o ambiente de TI local com o ambiente de Cloud Pública**: tendo a funcionalidade local semelhante a AWS torna mais fácil, barato e sustentável mover computação para AWS
- **Prover uma plataforma de desenvolvimento de software para a comunidade open source**

Arquitetura do Eucalyptus



Componentes do Eucalyptus

Cloud Controller – Componente "Base" Api - Amazon EC2;

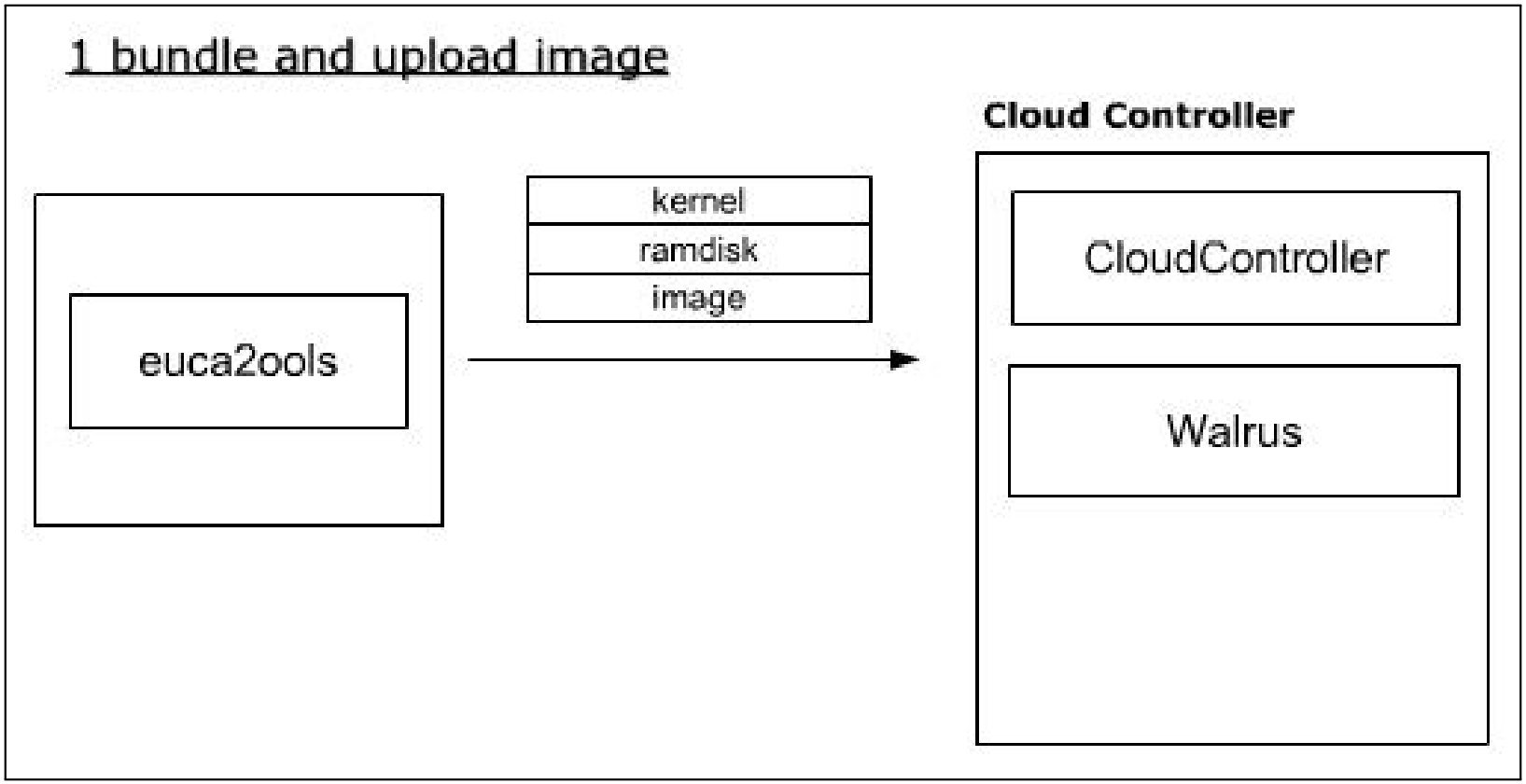
Walrus - "put-get bucket storage" Api – Amazon S3;

Cluster Controller – Controla os Node Controllers (VM, rede);

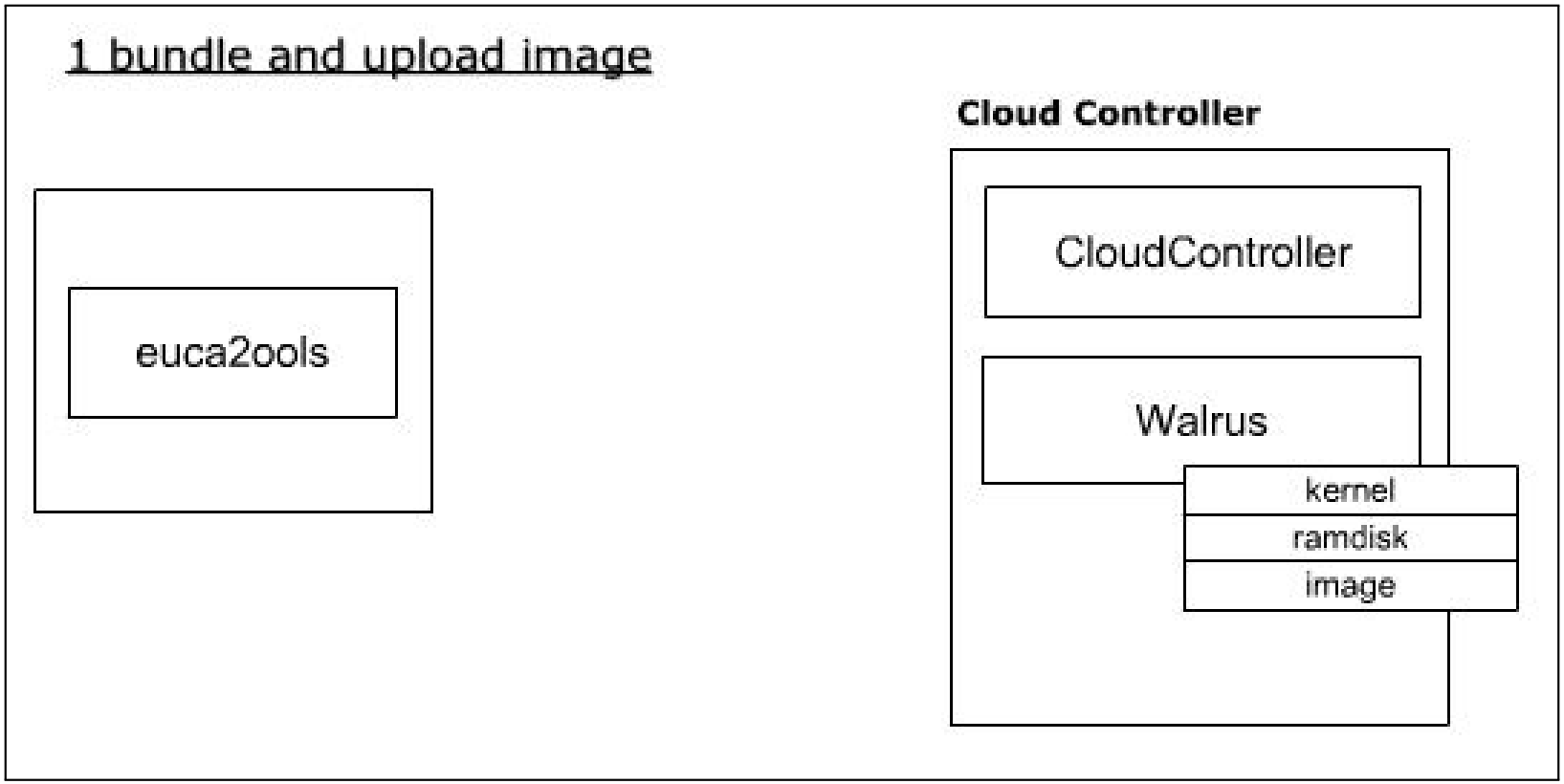
Storage Controller – Controla o "attach/detach" de um Volume a uma instância Api – Amazon EBS;

Node Controller – Componente que controla directamente o Hypervisor (execução, monitorização, finalização).

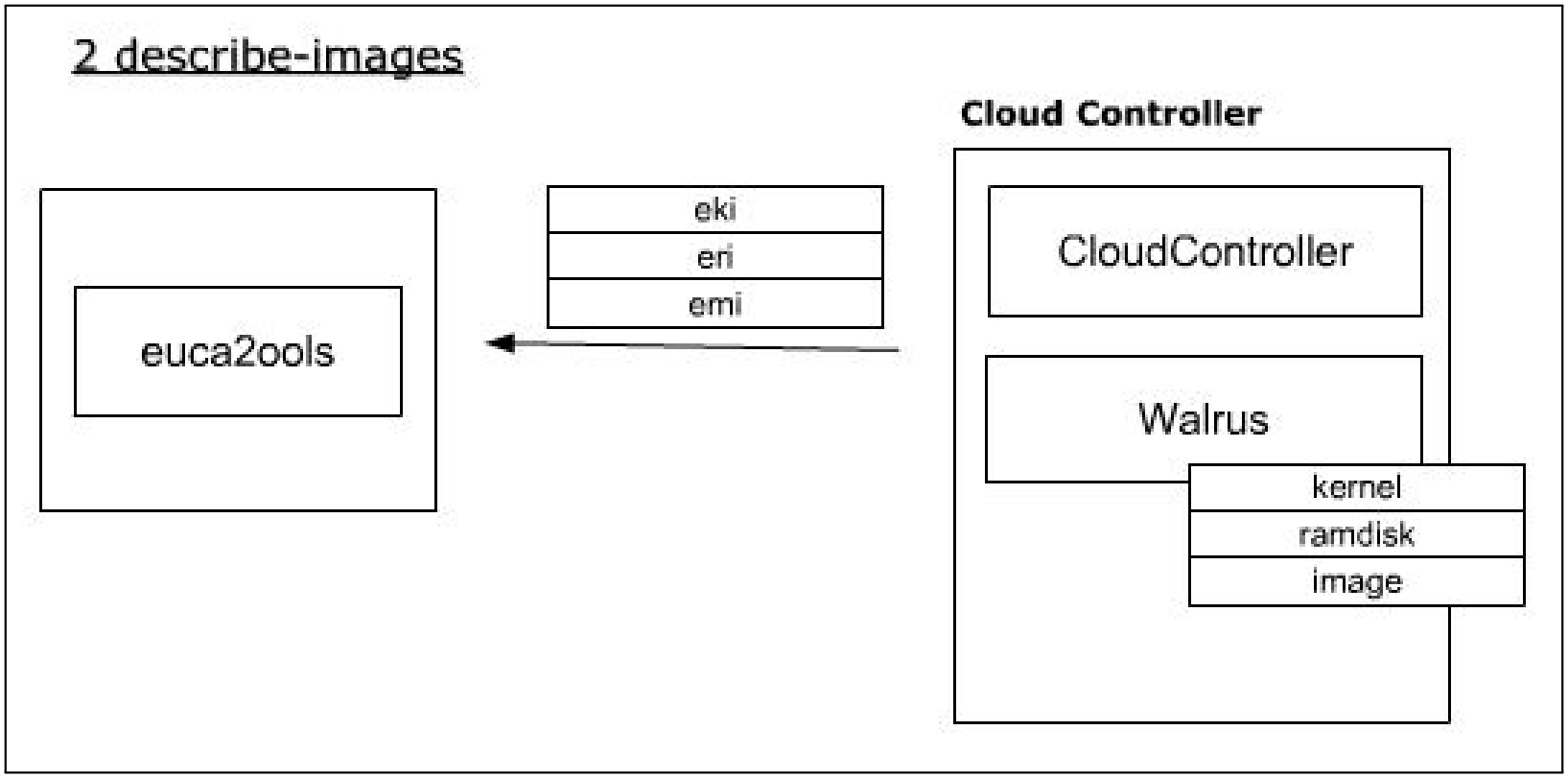
Funcionamento do Eucalyptus



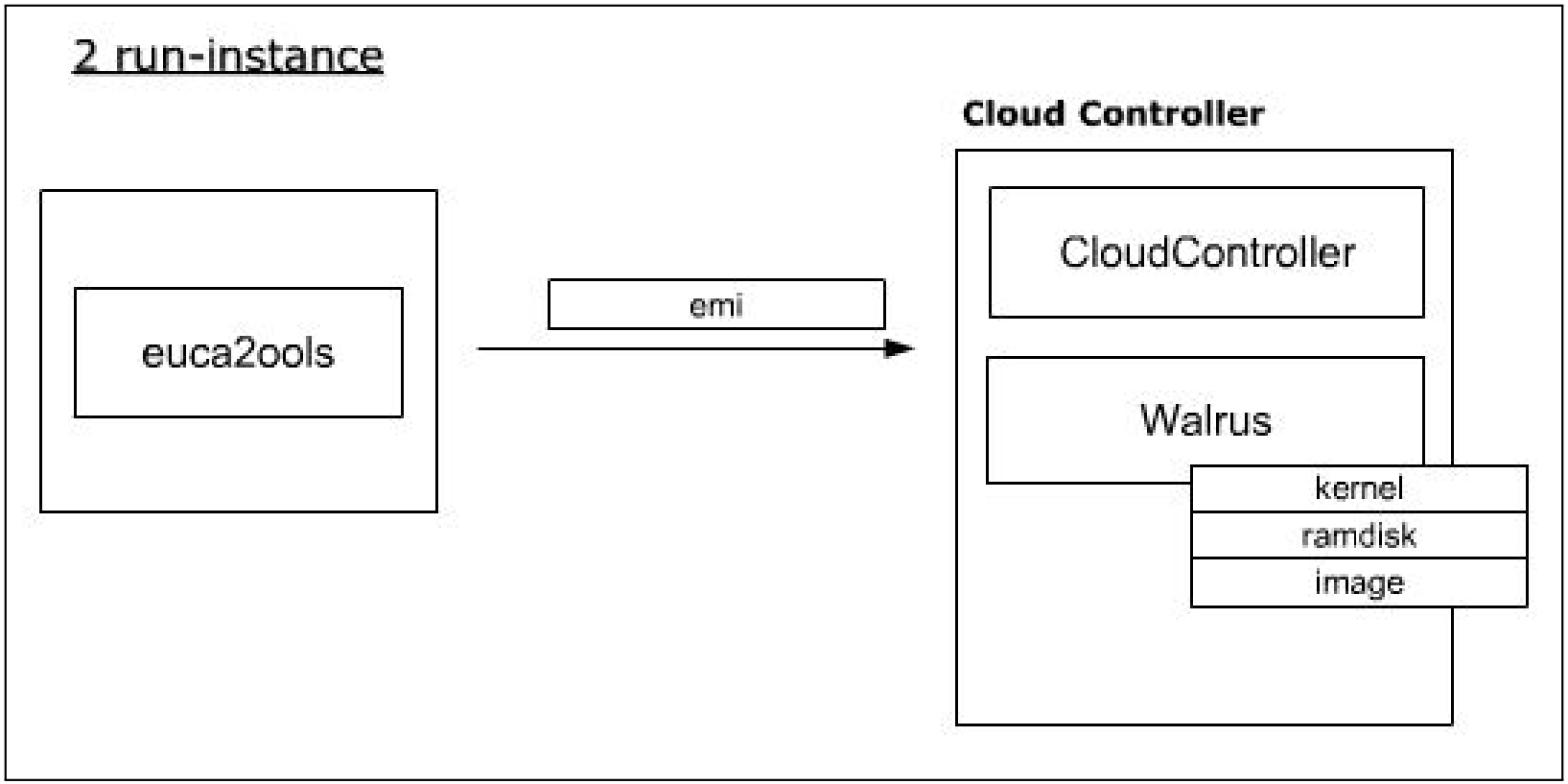
Funcionamento do Eucalyptus



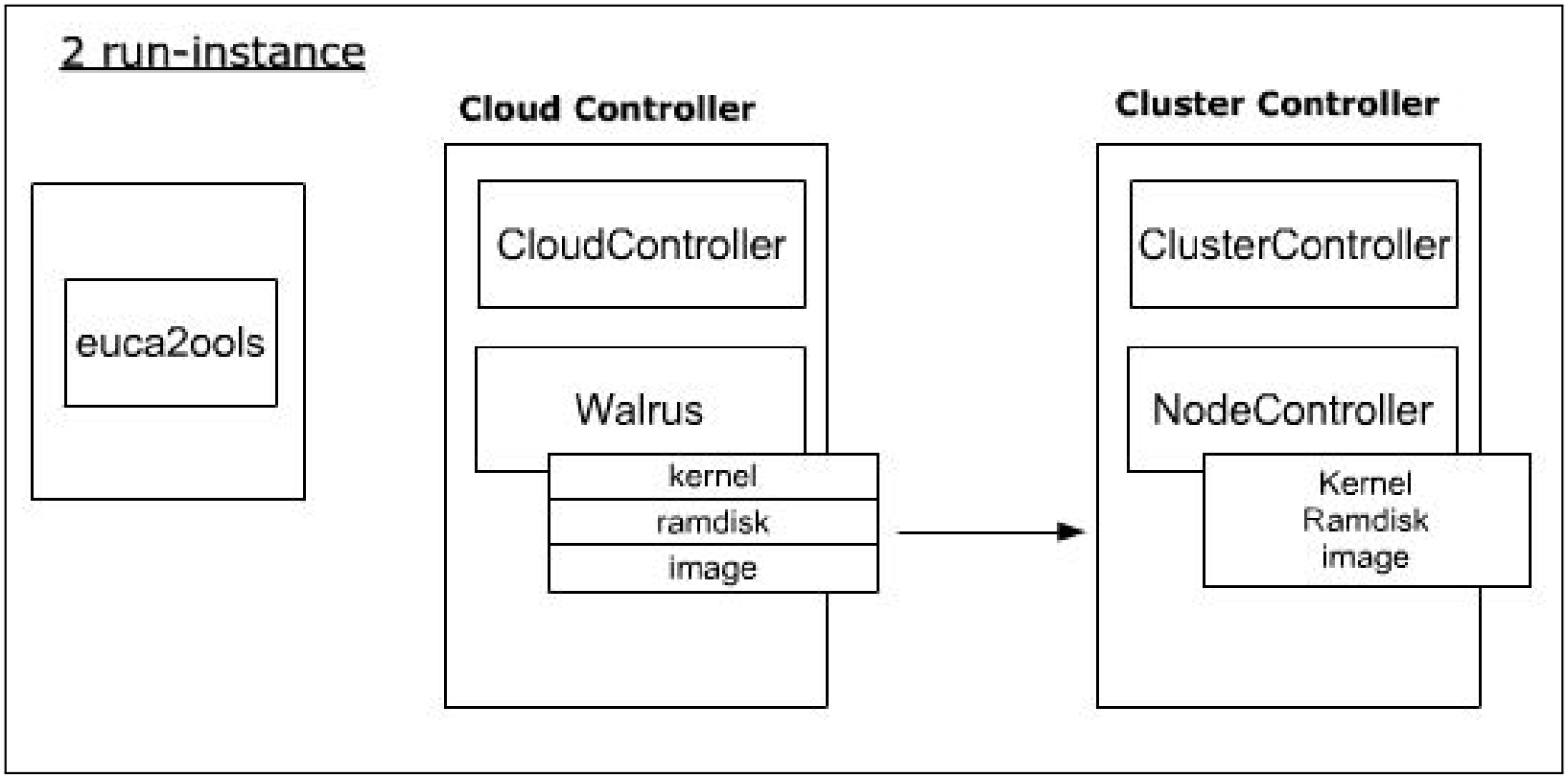
Funcionamento do Eucalyptus



Funcionamento do Eucalyptus



Funcionamento do Eucalyptus



Criando um Volume de Dados para Instancias - Eucalyptus

Create a 10 GB volume:

```
uecadmin@client1:~$ euca-create-volume -s 10 -z mycloud
VOLUME vol-333C04B8 10 creating
2010-03-26T05:20:56.383Z
```

List the volumes:

```
uecadmin@client1:~$ euca-describe-volumes
VOLUME vol-333C04B8 10 mycloud available
2010-03-26T05:20:56.383Z
```

Attach a volume to a running instance:

```
uecadmin@client1:~$ euca-attach-volume -i i-41620887 -d /dev/sdb
vol-333C04B8
```

Configurando Grupos de Segurança Avançados - Eucalyptus

Create a security group named “webservers”:

```
uecadmin@client1:~$ euca-add-group -d "Web Servers" webservers
```

Add a rule to the security group “webservers” allowing icmp and tcp traffic from a.b.c.d:

```
uecadmin@client1:~$ euca-authorize -P tcp -s a.b.c.d webservers
uecadmin@client1:~$ euca-authorize -P icmp -s a.b.c.d webservers
```

The added rules can be viewed with euca-describe-groups command:

```
uecadmin@client1:~$ euca-describe-groups
GROUP admin default default group
GROUP admin webservers Web Servers
PERMISSION admin webservers ALLOWS icmp -1 -1 FROMCIDR 0.0.0.0/0
PERMISSION admin webservers ALLOWS tcp 22 22 FROMCIDR 0.0.0.0/0
```

Eucalyptus

DEMO

Por Pedro Bignato (aluno graduação UFSCar)

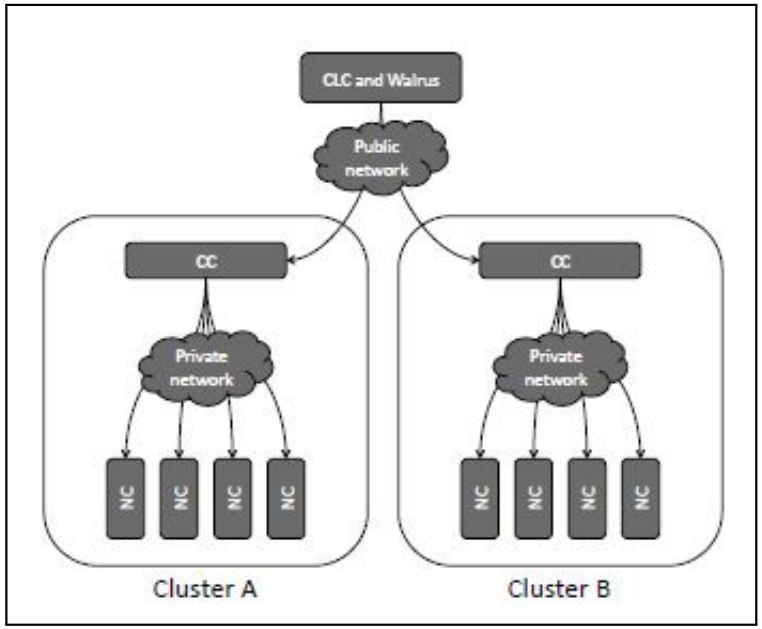
Eucalyptus - Experimentação com Cluster Privado

Lendas Urbanas sobre Computação em Nuvem

- Lenda 1: A Intraestrutura de Computação em Nuvem é só uma interface web para virtualização de sistema operacional
  - 'Estou executando Xen no meu datacenter, LOGO estou executando uma cloud privada'
- Lenda 2: Computação em Nuvem impõe uma penalidade significativa no desempenho comparado com o provisionamento 'bare metal'
  - 'Não poderei rodar uma cloud privada porque meus usuários não toleram uma queda de desempenho.'
- Lenda 3: Clouds e Grids são equivalentes
  - 'Nos anos 90, o termo grid foi criado para descrever tecnologias que permitiriam consumidores obter poder computacional sob demanda.'

Lenda 1: Cloud e Virtualização

- A virtualização de Sistema Operacional (Xen, KVM, VMWare, HyperV) é somente aparente no modelo IaaS, em outros ambientes como AppEngine existe mais camadas de software para elasticidade e auto-gerenciamento.
- Além disso, Hypervisors virtualizam CPU, memória e acesso a dispositivos locais como um ÚNICA máquina virtual (VM)
- A Alocação Cloud no modelo IaaS demanda mais que isso:
- Demanda a Alocação de VMs em recursos físicos
- Configuração e Uso de Conjuntos de Recursos de Armazenamento
- Configuração de Redes Privadas e Segurança



Lenda 2: Velocidade da Cloud - Eucalyptus

- Foram realizados extensos estudos de desempenho HPC
- Duas Questões Principais
  - Qual é o impacto de desempenho da virtualização?
  - Qual é o impacto de desempenho da infraestrutura de cloud?
- Os testes foram realizados com Xen e Eucalyptus, versus AWS (usando a instancias smalls)
- Muitas Respostas
  - Acesso Aleatório a Disco é mais lento com Xen
  - Aplicações CPU bound podem rodar mais rápido com Xen (paravirtualizado) do que com emulação de hardware
  - Algumas vezes a versão do Kernel, optimizações do compilador são mais importantes que o impacto da paravirtualização
  - O Eucalyptus não impõe como infraestrutura de cloud nenhum overhead estatisticamente importante
  - No caso do AWS, as instancias small aparentam ter limitações (throttle) de banda de rede - 0.10 / CPU hour

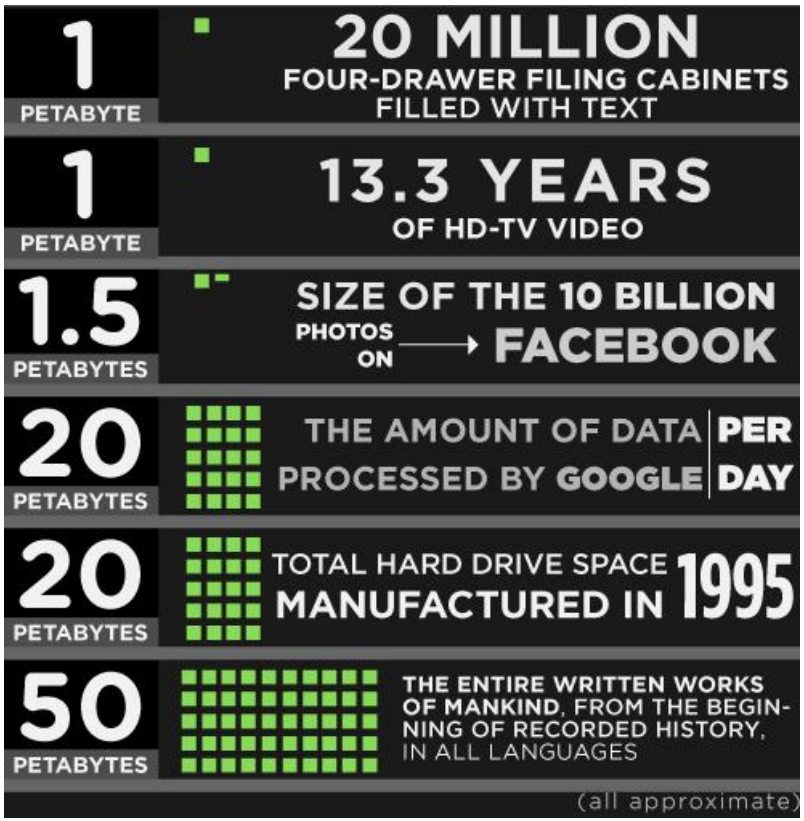
Lenda 3: Clouds versus Grid

- Cloud
  - Cluster Privado Completo é Provisionado
  - Usuário individual obtém uma pequena porção do pool de recursos totais
  - Nenhum suporte a federação de cloud exceto através da interface cliente
  - Opaco com relação a recursos
- Grids
  - Por design feito de modo que usuários individuais tenham acesso a maioria, se não todos, os recursos com uma única requisição
  - Abordagem com Middleware usa a federação como um primeira abordagem
  - Recursos são expostos, frequentemente como bare metal
- Estas diferenças requerem diferentes arquiteturas para cada: cloud vs grid

Tendencias em Processamento de Dados de Alto Desempenho

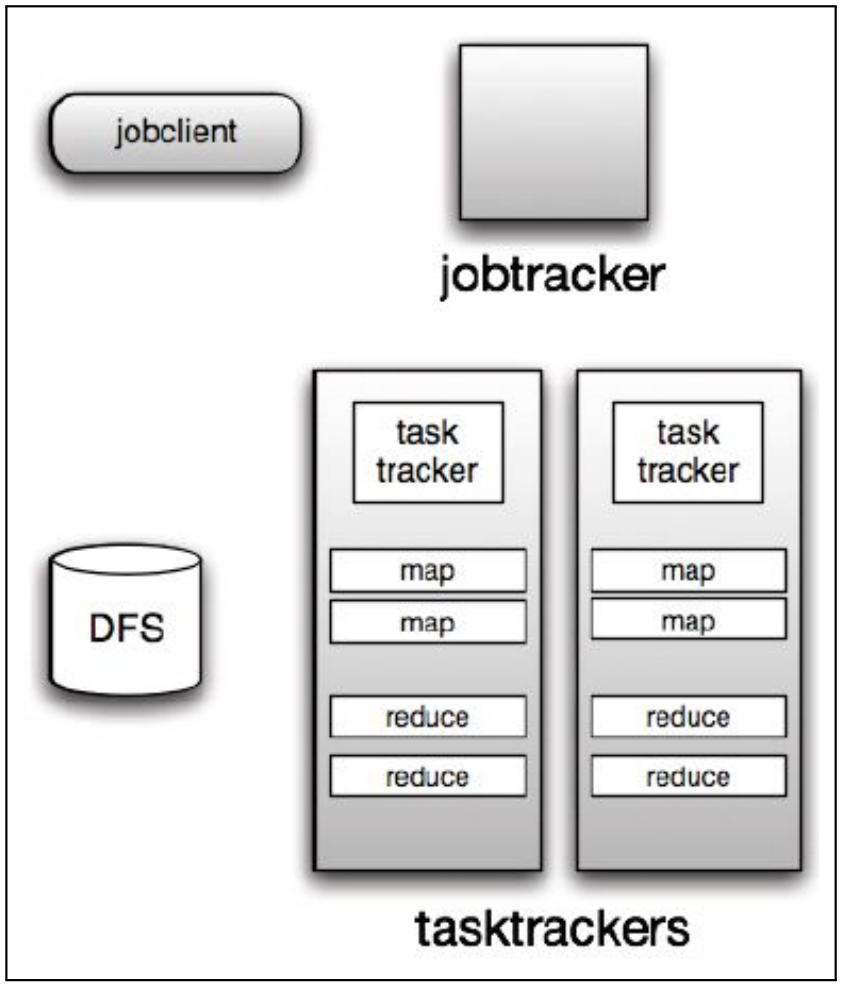
- NYSE gera cerca de 1 Terabyte de dados por dia
- LHC irá produzir 15 Petabytes de dados por ano
- Pense na quantidade de dados presentes em tweets, facebook, fotos, comentários em slashdot - estima-se 1 Tb/ano por pessoa
- Onde armazenar? Como ler? Como processar?

RESPOSTA: WORKFLOWS

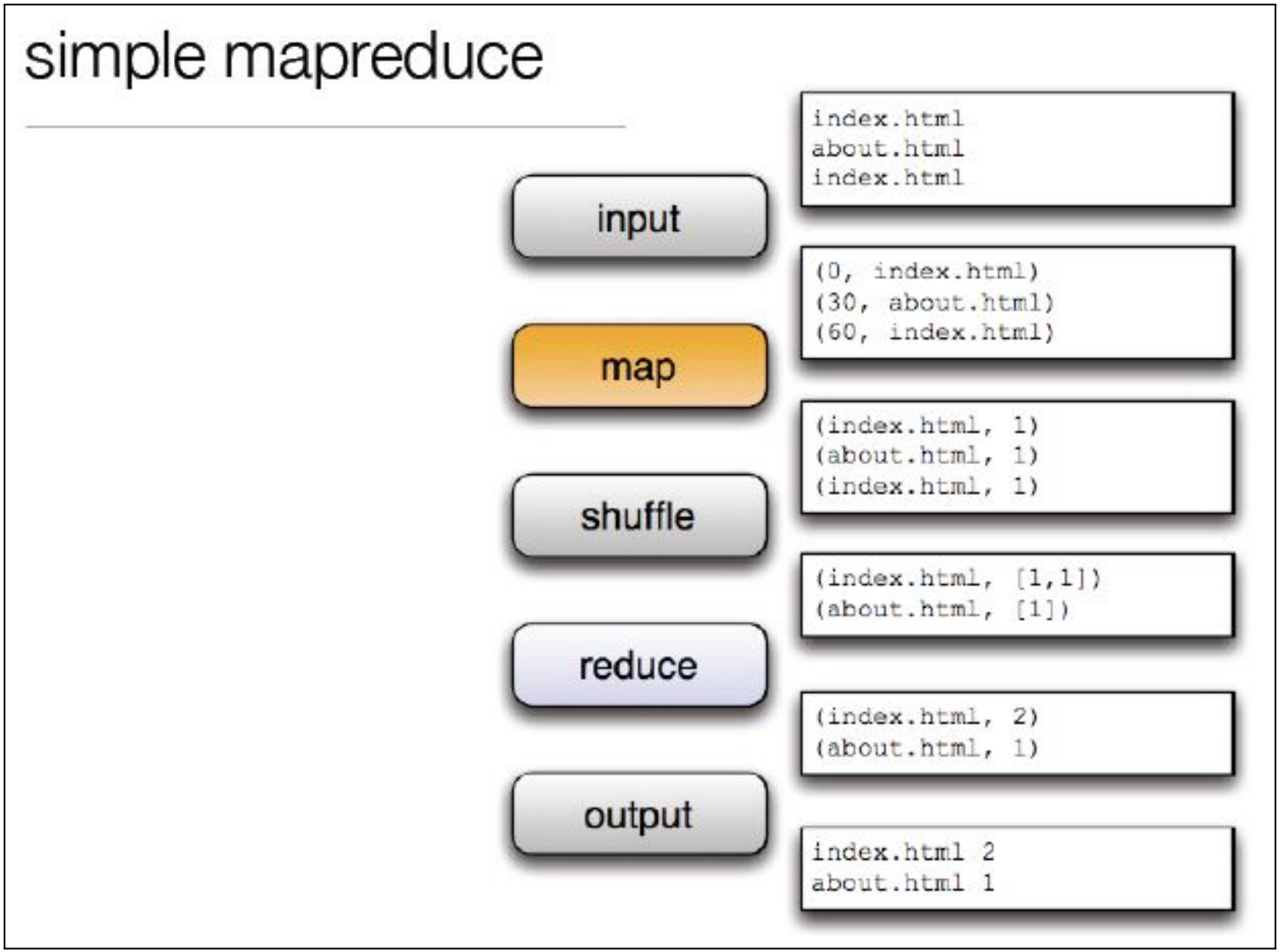


Introduzindo MapReduce

- criado por Jeff Dean e Sanjay Ghemawat (google), baseia-se em funções de programação funcional (Lisp) ’map’ e ’reduce’
- distribui a carga usando mestre/escravos e lê/escreve de um sistema de arquivos distribuído
- **Existem 3 regras a seguir sobre como paralelizar grandes bases de dados. Infelizmente, ninguém sabe quais regras são estas - Gary R. Montry**

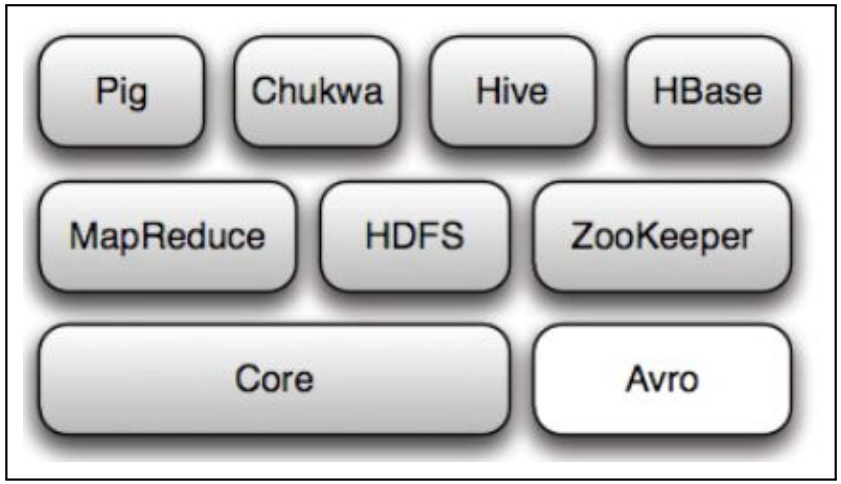


MapReduce Simples



Apache Hadoop

- Implementação MapReduce mais difundida
- Projeto considerado top level do apache project desde jan 2008
- Código-aberto, baseado em java
- Vencedor do benchmark terabyte sort
- Investimentos pesados pela Yahoo! e usado em seus clusters



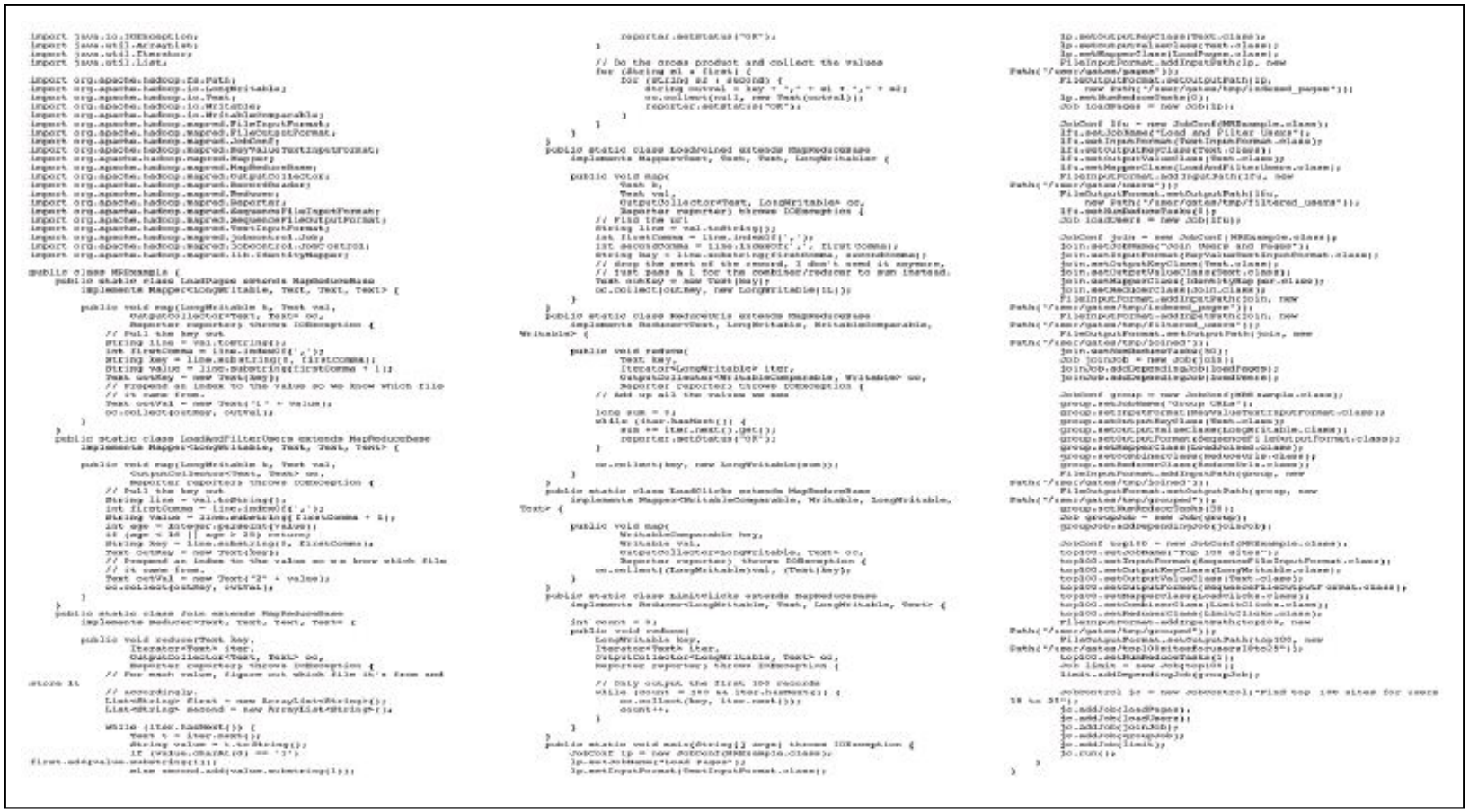
Pig fazendo o Resgate da Complexidade

- torna mais simples escrever programas mapreduce - parecido com um SQL
- PigLatin abstrai do usuário os detalhes específicos e foca no processamento de dados

Um Exemplo Complexo de Processamento de Dados

- Dados de usuários (user data) em um único arquivo
- Visitas ao website espalhados em dúzias de logs
- Encontrar as top 5 páginas mais visitadas por usuários com idade entre 18 e 25 anos

Escrevendo o Código em MapReduce/Hadoop



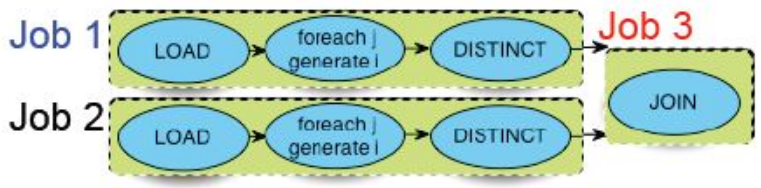
E Agora Comparado ao Pig

```
Users = LOAD 'users' AS (name, age);
Fltrd = FILTER Users BY
        age >= 18 AND age <= 25;
Pages = LOAD 'pages' AS (user, url);
Jnd = JOIN Fltrd BY name, Pages BY user;
Grpd = GROUP Jnd BY url;
Smmd = FOREACH Grpd GENERATE group,
       COUNT(Jnd) AS clicks;
Srtd = ORDER Smmd BY clicks DESC;
Top5 = LIMIT Srtd 5;
STORE Top5 INTO 'top5sites';
```

Transformação

Pig Latin Script → MapReduce DAG

```
A= LOAD 'file1' AS(user,action);
B= FOREACH A GENERATE user;
C= DISTINCT B;
D= LOAD 'file2' AS(user,action);
E= FOREACH D GENERATE user;
F= DISTINCT E;
G= JOIN C BY user,F BY user;
   STORE G into 'output.txt';
```



Automatizando Hadoop Workflow no Cloud AWS

- Amazon Elastic MapReduce é um serviço web que permite a empresários, pesquisadores, analistas de dados e desenvolvedores a facilmente de uma maneira eficiente do ponto de vista de custo processar vastas quantidades de dados

- Esse serviço usa um framework Hadoop nativo hospedado no AWS rodando na infraestrutura de escala-web da Amazon Elastic Compute Cloud (Amazon EC2) e da Amazon Simple Storage Service (Amazon S3).

Region: US East (Virginia)

	Amazon EC2 Price	Amazon Elastic MapReduce Price
Standard On-Demand Instances		
Small (Default)	$0.085 per hour	$0.015 per hour
Large	$0.34 per hour	$0.06 per hour
Extra Large	$0.68 per hour	$0.12 per hour

Contabilização é Complicada - Extraído do FAQ

- Contabilização começa quando o seu job flow Amazon Elastic MapReduce (e quando é isso?)
- Por exemplo, digamos que você inicie 100 Amazon EC2 Standard Small instances para um job flow Amazon Elastic MapReduce
- As instancias Amazon EC2 irão começar a bootar imediatamente, mas não necessariamente começarão no mesmo momento
- De qualquer forma, o serviço Amazon Elastic MapReduce irá rastrear cada instância que começar e fará a marcação dela no cluster de modo a aceitar tarefas

Contabilização é Complicada - Extraído do FAQ - Parte 2

- Nos primeiros 10 minutos após ter feito a requisição do serviço Amazon Elastic MapReduce
- Por exemplo, digamos que você inicie 100 Amazon EC2 Standard Small instances para um job flow Amazon Elastic MapReduce
- OU o job flow inicia (se todos as suas instancias estiverem disponíveis)
- OU faz o check in de quantas instancias forem possíveis. E uma vez que passarem 10 minutos cravados, Amazon Elastic MapReduce começará o processamento (e a cobrança) assim que 90% das instancias requisitadas do seu job flow estiverem disponíveis. Assim que o restante 10% das instancias requisitadas fizerem check-in Amazon Elastic MapReduce começara a cobrar por estas instancias também
- No exemplo acima, se todas as 100 das instancias requisitadas estiverem disponíveis nos 10 minutos após a requisição de launch, voce será cobrado $1.50 por hora (100 * $0.015) até o job flow completar. E assim sucessivamente para o caso 90%.

Amazon Elastic MapReduce

DEMO

Por Amazon AWS usando PIG

http://s3.amazonaws.com/awsVideos/AmazonElasticMapReduce/ElasticMapRe
PigTutorial.html

WorkFlows de Processamento de Dados Mais Agnósticos

- Embora o Hadoop já tenha sido incorporado em workflow bem específicos na Amazon AWS
- Existe a necessidade de ferramentas para processamento em lote mais agnósticas ao Cloud Provider
  - Apache Whirr
  - jClouds

O que é o Projeto Apache Whirr?

- Um conjunto de bibliotecas para rodar serviços na nuvem
- Cloud-neutral
- API Comum de Serviço
- Provê configurações inteligentes
  - O código do Whirr começou em 2007 como um conjunto de scripts bash para executar Apache Hadoop em clusters da Amazon EC2
  - Mais tarde estes scripts foram portados para Python para features extras (como suporte a EBS) e vários cloud providers
  - Em 2010 foi escolhida a API do jclouds que suporta uma dezena de provedores e tem uma API rica para rodar código em instancias, de modo que provê uma base sólida para a construção do Whirr.

O que se pode fazer com o Apache Whirr?

- Instalar e Configurar Clusters sob demanda para processamento ou para testes (injeção de falha, testes de regressão). Ideal quando desenvolvendo com aplicações no topo de componentes da stack Hadoop
- Orquestra clusters e aplicações baseado em papéis e configurações
- Serviços Suportados: Cassandra, Hadoop, Hbase, ZooKeeper
- Provedores de Cloud Supportados: EC2, RackspaceCloud (usando jclouds)

Configuração Típica do Whirr

whirr.cluster-name=hadoop

whirr.instance-templates=1 hadoop-namenode+hadoop-jobtracker,5 hadoop-datanode+hadoop-tasktracker

whirr.provider=ec2

whirr.identity=AWS_ACCESS_KEY_ID

whirr.credential=AWS_SECRET

whirr.hardware-id=c1.xlarge

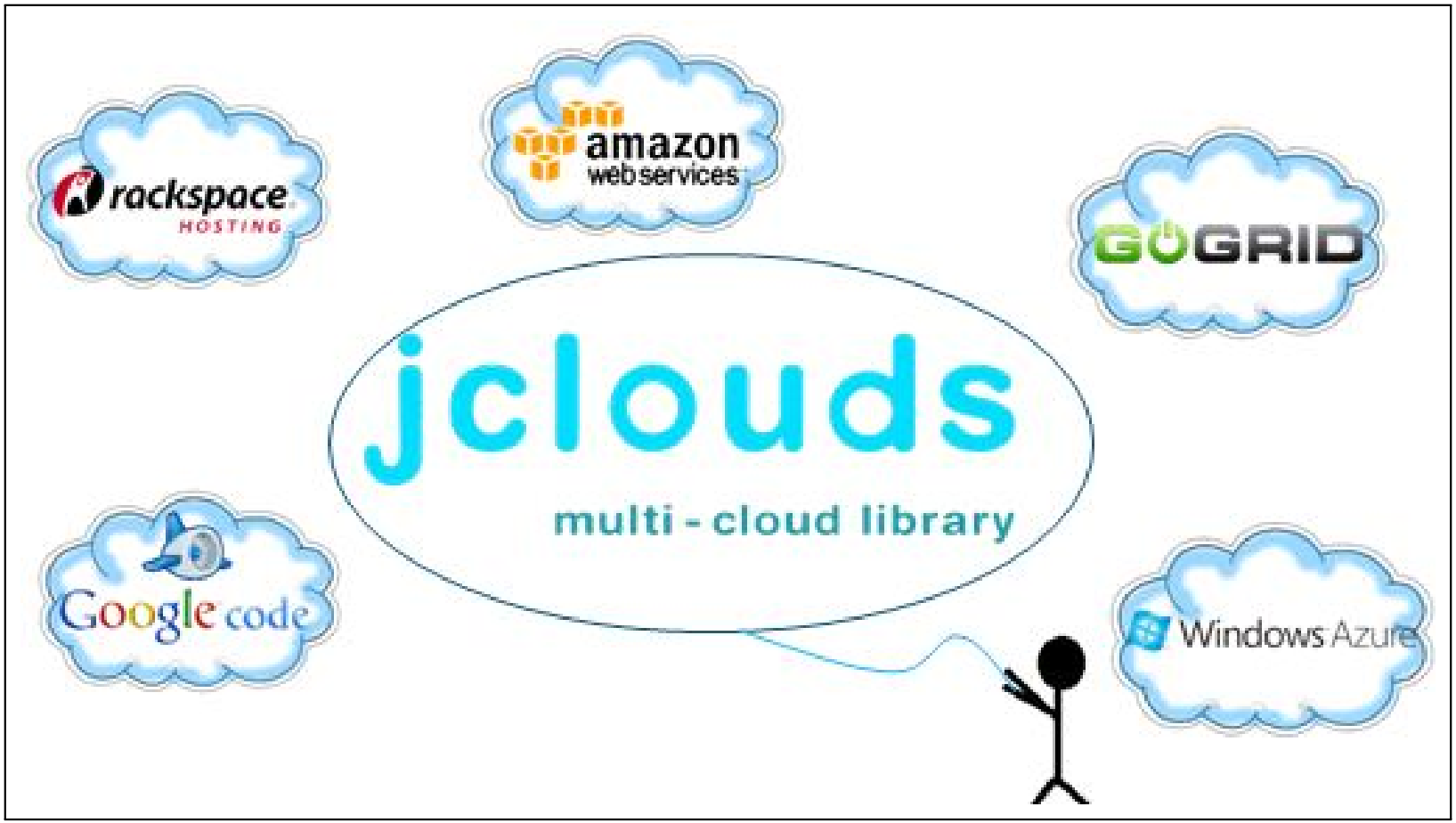
Executando Cluster Cassandra com Whirr

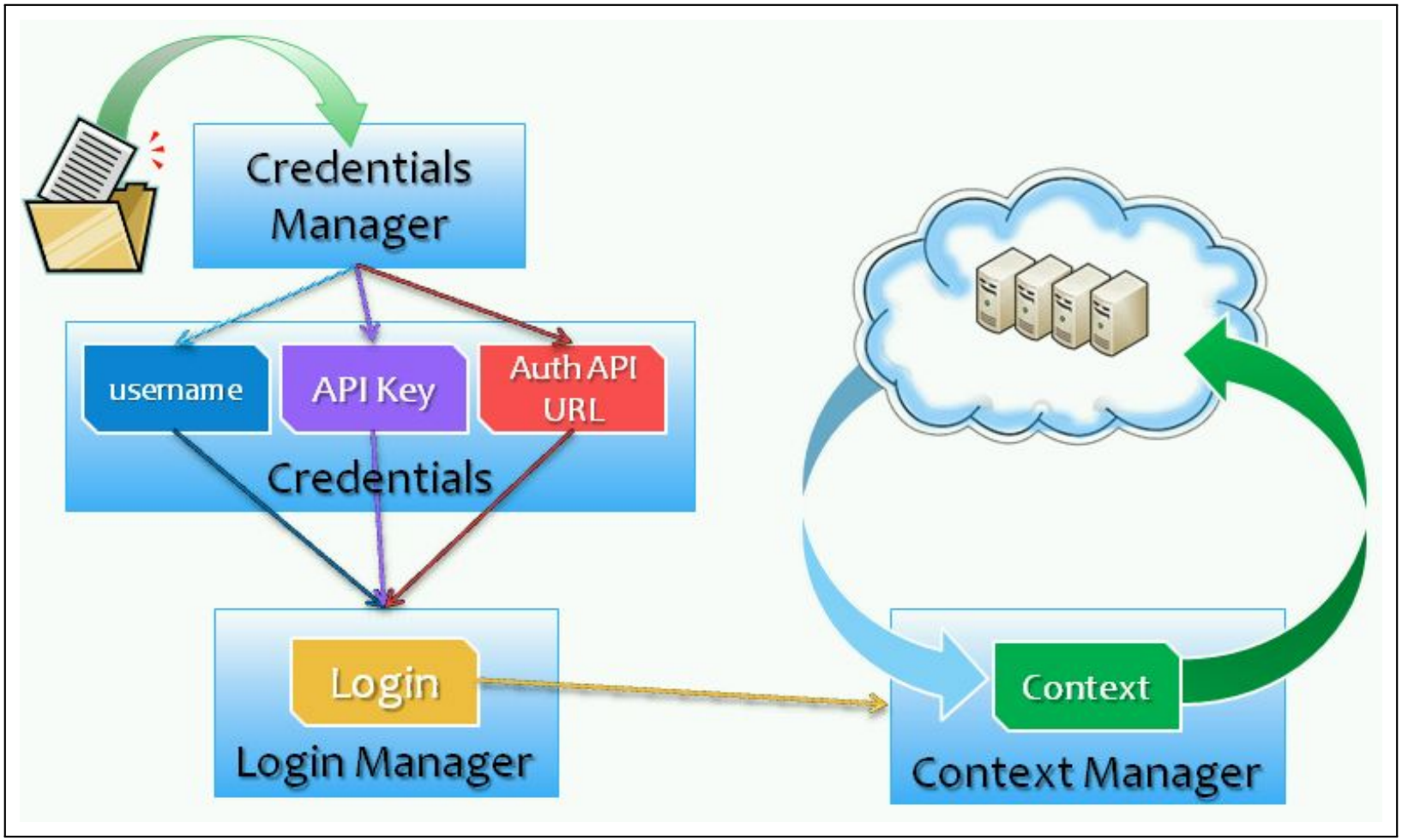
```
whirr.service-name=cassandra
whirr.cluster-name=mycassandracluster
whirr.instance-templates=3 cassandra
whirr.provider=ec2
whirr.identity=<YOUR_AMAZON_EC2_ACCESS_KEY_ID_GOES_HERE>
whirr.credential=<YOUR_AMAZON_EC2_SECRET_ACCESS_KEY_GOES_HERE>
whirr.private-key-file=${sys:user.home}/.ssh/id_rsa
```

```
bin/whirr launch-cluster --config cassandra.properties


Launching mycassandracluster cluster
Configuring template
Starting 3 node(s)
Nodes started: [[id=us-east-1/i-13f25e7f, providerId=i-13f25e7f, tag=mycassandracluster, name=null, loc
Authorizing firewall
Running configuration script
Completed launch of mycassandracluster
Started cluster of 3 instances
Cluster{instances=[Instance{roles=[cassandra], publicAddress=/50.16.85.79, privateAddress=/10.117.43.12
```

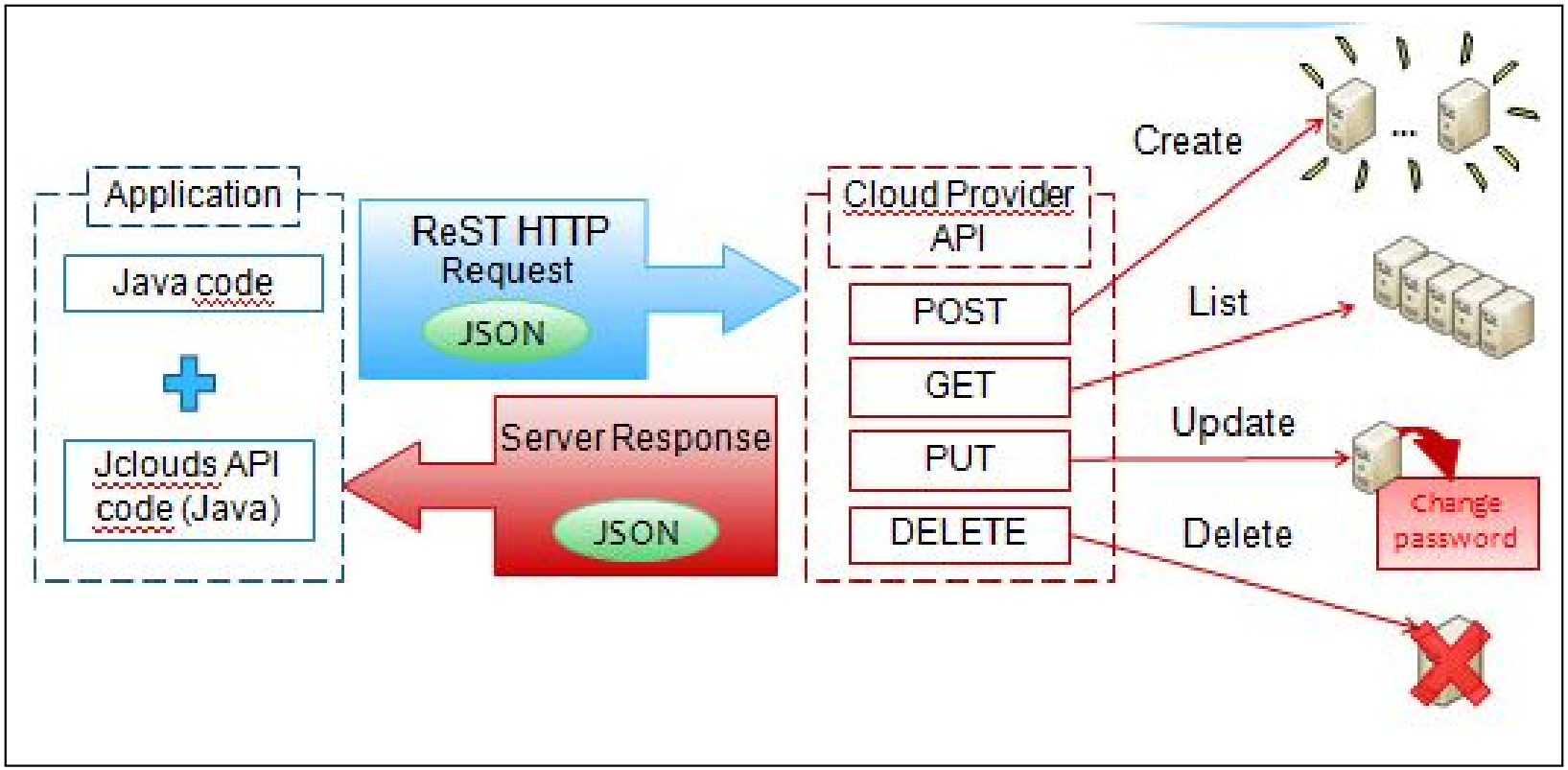
```
bin/whirr destroy-cluster --config cassandra.properties

Destroying mycassandracluster cluster
Cluster mycassandracluster destroyed
```

jClouds


rackspace
HOSTING
amazon
web services
GOGRID
jclouds
multi-cloud library
Google code
Windows Azure


Credentials Manager
username
API Key
Auth API URL
Credentials
Login
Login Manager
Context
Context Manager

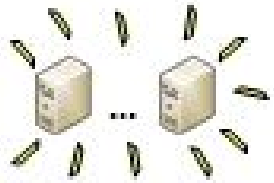
jClouds


Application
Java code
Jclouds API code (Java)
ReST HTTP Request
JSON
Server Response
JSON
Cloud Provider API
POST
GET
PUT
DELETE
Create
List
Update
Delete
Change password

jClouds

```
ComputeServiceContext context =
new ComputeServiceContextFactory().createContext(
"cloudservers-us", wiring, overrides
);


CloudServersClient rackspaceClient =
CloudServersClient.class.cast(
context.getProviderSpecificContext().getApi()
);

BaseComputeService client =
BaseComputeService.class.cast(context.getComputeService());
```

```
Template nodeToCreate =
context.getClient().templateBuilder().osFamily(OsFamily.UBUNTU)
.osVersionMatches("10.10").minCores(2).minRam(minRam).options(options
).build();

System.out.println("Creating nodes ...");

Set<? extends NodeMetadata> nodes =
context.getClient().createNodesInGroup(nodeGroupName, numberOfNodes,
nodeToCreate);
```



jClouds

```
context.getClient().rebootNode(node);

context.getRackspaceClient().resizeServer(serverid, 3);

context.getClient().destroyNode(nodeToDelete);
```

jClouds

DEMO

Por Bruno Katekawa (aluno graduação UFSCar)

jClouds para Computação de Alto Desempenho

Workloads Ciêntificos na Computação em Nuvem

- Trata-se uma Área Nova e Promissora para Uso Esporádico
- Conferencias Específicas como ScienceCloud 2010 e ScienceCloud 2011
- Foram Feitas Extensas Avaliações de Desempenho em Cloud Providers
- Farei um Pequeno Resumo de Alguns dos Mais Relevantes Trabalhos
- Ao Final - Recursos Recentes da AWS e DEMO

Cloud Computing for parallel Scientific HPC Applications - Evangelinos - MIT 2008

- Será que a Computação em Nuvem pode ser o caminho para HPC científica?
  - Ela pode ser útil para mais do que tipos de aplicações EP ou Map-Reduce?
  - Será que os dias de ter que comprar, instalar e manter clusters pessoais está chegando ao fim?
  - Seria possível usar dinheiro de pesquisa (grants) para comprar ciclos de computação em nuvem?
  - Seria possível usar Computação em Nuvem para ensino de HPC?
- E sobre o desempenho HPC em um ambiente de máquina virtual?
  - Qual máquina, software, combinação funciona melhor?
  - Poderia ser usado nossos próprios clusters?

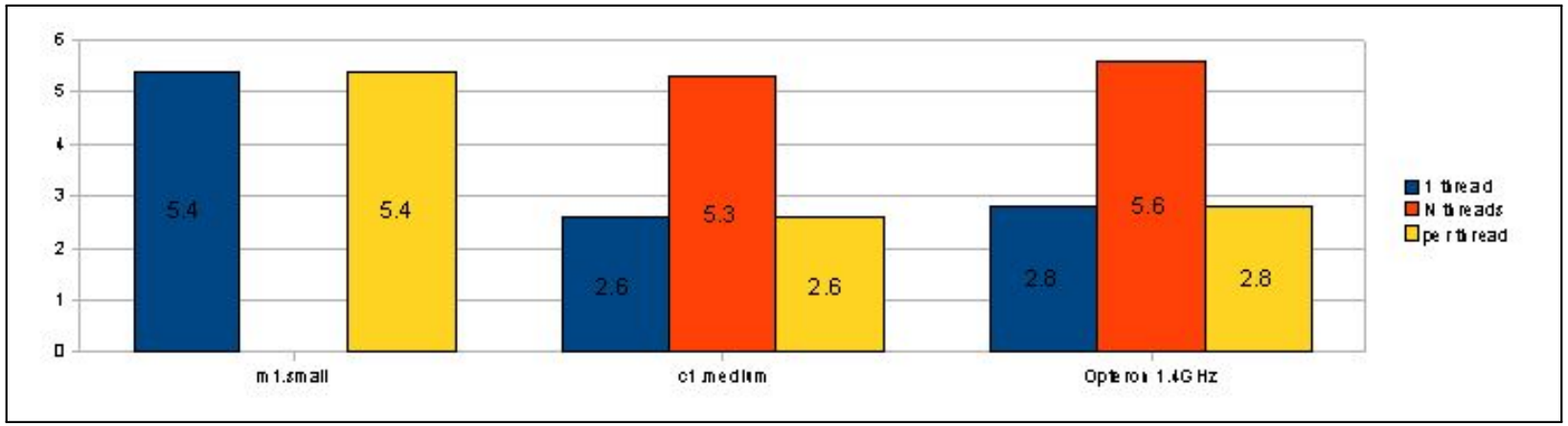
Instâncias Amazon EC2 - Evangelinos - MIT 2008

'UMA Unidade Compute EC2 equivalente a capacidade CPU 1.0-1.2 GHz de processador Opteron ou Xeon 2007.'

- Instancias Standard: suficientes para a maioria das aplicações
- m1.small - Small Instance (default): 1.7 GB memory 1 EC2 Compute Unit (1 virtual core with 1 EC2 Compute Unit), half of a core of a 2 socket node, dual core AMD Opteron(tm) Processor 2218 HE, 2.6GHz, 160 GB instance storage (150 GB plus 10GB root partition), 32-bit platform, moderate I/O performance,price: $0.10 per instance hour
- m1.large - Large Instance: 7.5 GB memory, 4 EC2 Compute Units (2 virtual cores with 2 EC2 Compute Units each), half of a 2 socket node, dual core AMD Opteron(tm) Processor 270, 2.0GHz, 850 GB instance storage (2 x 420 GB plus 10 GB root partition), 64-bit platform, high I/O performance, price: $0.40 per instance hour
- m1.xlarge - Extra Large Instance: 15 GB memory, 8 EC2 Compute Units (4 virtual cores with 2 EC2 Compute Units each), one 2 socket node, dual core AMD Opteron(tm) Processor 270, 2.0GHz, 1,690 GB instance storage (4 x 420 GB plus 10 GB root partition), 64-bit platform, high I/O performance, price: $0.80 per instance hour

- Instancias High-CPU: menor RAM por core, apropriado para aplicações CPU-intensive
- c1.medium - High-CPU Medium Instance: 1.7 GB of memory, 5 EC2 Compute Units (2 virtual cores with 2.5 EC2 Compute Units each), half of a 2 socket node, quad core Xeon E5345, 2.33GHz, 350 GB of instance storage, 32-bit platform, moderate I/O performance, price: $0.20 per instance hour
- c1.xlarge - High-CPU Extra Large Instance: 7 GB of memory, 20 EC2 Compute Units (8 virtual cores with 2.5 EC2 Compute Units each), one 2 socket node, quad core Xeon E5345, 2.33GHz, 1690 GB of instance storage, 64-bit platform, high I/O performance, price: $0.80 per instance hour

Impacto Banda de Memória - Evangelinos - MIT 2008



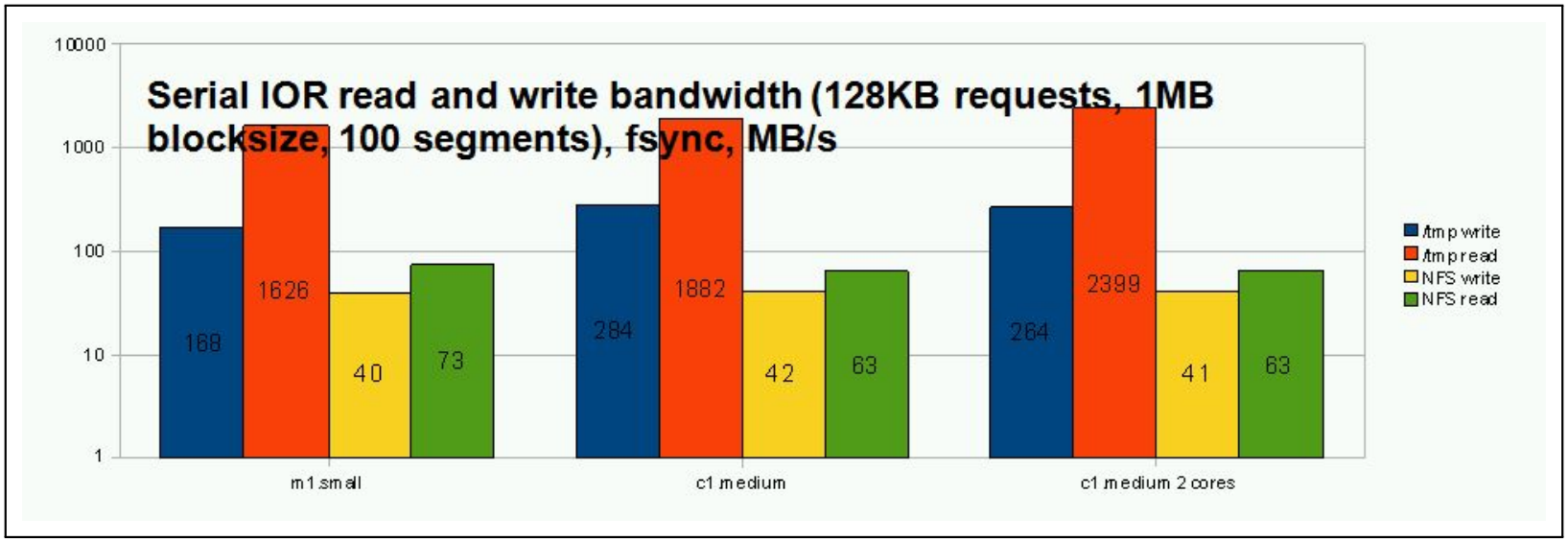
- A largura de banda de memória para a 'small instance' aparenta ser equivalente ao total de memória esperado de tal plataforma. Embora, exista um throttler de 50% CPU. Comparado com um computador de 2007 (1.4 GHz sistema Opteron) também foi pior. A memória DDR2 deve ajudar.
- Uma CPU mais rápida usando instância c1.medium teve resultado consideravelmente pior. Embora, depois o aumento da largura de banda de memória com a instância c1.medium sugere que os dois núcleos não estão no mesmo chip.

HPC Benchmark - Evangelinos - MIT 2008

System	Class A	Class W	EP (A)	EP (W)
m1.small	132	149	6.66	6.73
c1.medium	312	357	15.59	15.04
ratio	2.36	2.4	2.34	2.23

- Versão Serial do NAS Parallel Benchmark NPB (usando média geométrica com todos os testes exceto EP)
- Resultados são em Mop/s ou Milhões Numeros Aleatórios/s - STREAM benchmark
- Parallel Class (A) e Workstation Class (W) especifico do NPB
- Compilado com sistema de gcc (flags genérico)
- Única instância em execução no caso c1.medium (sem contenção de memória)
- Proporção Teórica 1:5, Proporção Obtida 1:2.3. Relação de Preço 1:2

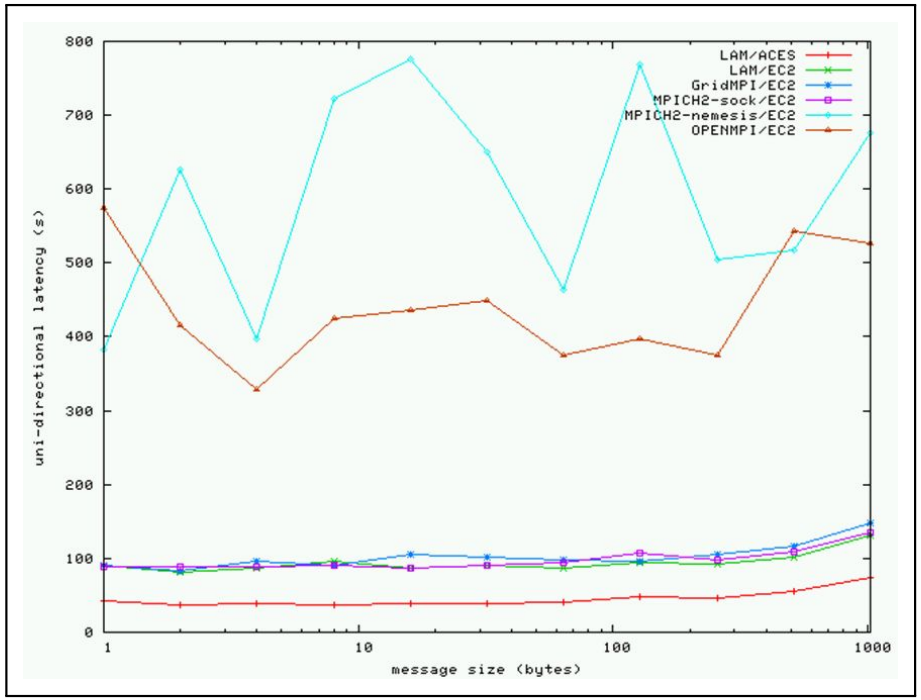
Desempenho de I/O - Evangelinos - MIT 2008



- Diferença sensível entre a escrita e o desempenho de leitura usando instância small e high-cpu.
- Desempenho NFS é essencialmente o mesmo. Todos em MB/s - IOR benchmark
- Usando ambos os núcleos c1.medium aumenta leitura e piora escrita.
- Em termos de I/O usar a opção c1.medium parece ser a melhor opção.
- Novamente desempenho não é 5 vezes melhor, mas fica entre 1.5 e 2.0. e preço é 2.0

MPI em EC2 - Evangelinos - MIT 2008

- Utilizado MPI benchmarks (OMB v. 3.1) e instâncias m1.small.
- Expectativas:
  - MPICH2 nemesis e OpenMPI melhor desempenho
  - LAM (Open Cluster MPI) ter desempenho respeitável
  - Idem para MPICH2 sockes
  - GridMPI (sendo orientados a WAN) teria o pior desempenho
- A dura realidade:
  - MPICH2 nemesis e OpenMPI levam séculos para terminar devido a problemas de latencia e largura de banda.
  - LAM tem o melhor desempenho!
  - Comparado com um cluster conectado a GigE ainda um fator de 2 a diferença em termos de latência e largura de banda.
- Resultado da Virtualização de Rede Xen + MPI -> futuro Xen-aware MPI



Cálculo de Custos - - Evangelinos - MIT 2008

- Parte de um exercício sobre necessidades futuras de cloud pelo MIT:
  - 0.8 cents/hora para instancia Xeon 2 cpu, 8-core no Amazon EC2 (opção mais barata com o máximo de flexibilidade disponível)
  - Custo de 158 racks, baixa densidade, 2U, 21 nós por rack equivale a 158x21x0.8 ou 2654,4 dólares por hora.
  - Assumindo um aproveitamento de 85%, isso equivale a 2654,4 * 24 * 365 * 0,85 = $ 19,8 milhões por ano, ou 8 vezes a nossa fatura do MIT de electricidade esperado para um datacenter altamente eficiente.
  - Mesmo com o custo de construção de datacenter nuvem teria um custo maior após 4 anos de uso.
- ENTRETANTO o uso esporádico é bastante adequado economicamente para o uso de nuvens.
  - 20 limitação instância do EC2 poderia ser um problema, precisa preencher formulário.
  - Limitações Gigabit Ethernet podem limitar o uso de grande número de instancias do tipo large.

Can Cloud Computing Reach The TOP500 - Napper - 2008

- Clouds provê um ambiente amplo e rapidamente escálavel de conjuntos de máquinas
  - Centenas de nós facilmente disponíveis
  - Maneira Similar a um cluster Beowulf no aspecto de rede 1GigE
- Entretanto, não há garantia de desempenho
  - Maquinas Virtuais podem reservar uma certa quantidade de RAM ou disco, mas não o acesso exclusivo a rede ou escalonamento exclusivo da CPU!
- Será possível termos a Computação em Nuvem alcançar a lista dos Top-500?

Ambiente e Configuração dos Testes - Napper - 2008

- Foi usado para testes de desempenho a implementação HPL do Linpack
- Resultados mostram que a memória e a banda de rede é insuficiente para manter alta performance ao aumentar a escalabilidade do cluster
- Algoritmo de Complexidade Cúbica para Crescimento Quadrático de Dados
- Uso Implementação MPI - MPICH e biblioteca BLAST
- Compilou com opções de otimização e número de threads específicos das instancias AMD e Intel
- Instancia high-cpu mapeada em Intel Xeon com 2.3Gh/7G com 4 fp/clock, instancia extra-large mapeada em AMD 2.6Gh/15GB com 2 fp/clock
- Linux 2.6.21, RedHat Fedora Core 8

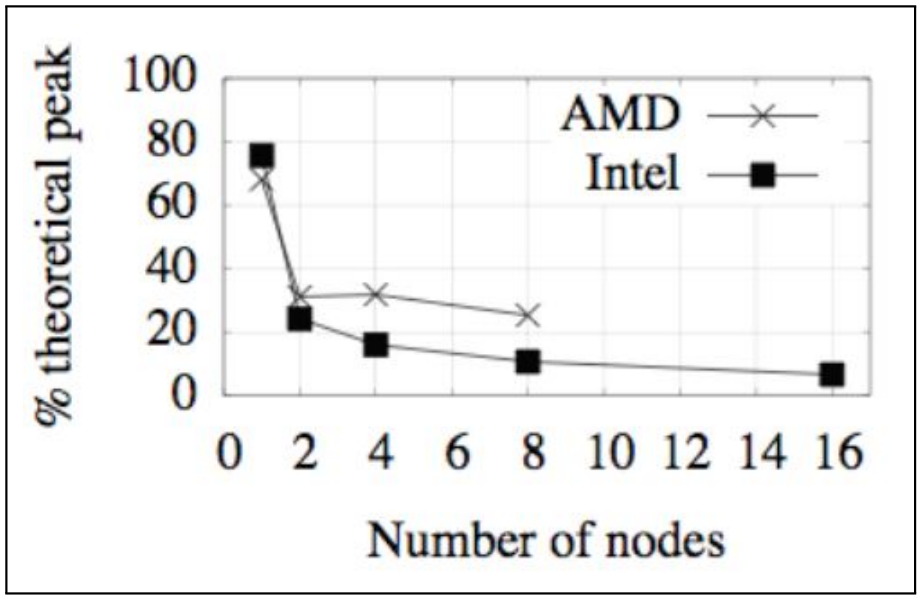
Instancias EC2 Usadas nos Experimentos - Napper - 2008

AMD: $0.80 / hour
- 2 X 2.6 GHz Dual-core Opteron
- 15 GB memory: 3.8 GB / core
- BLAS: 14 GFLOP/sec (68% theoretical peak)

Intel: $0.80 / hour
- 2 X 2.3 GHz Quad-core Xeon
- 7 GB memory: 0.9 GB / core
- BLAS: 57 GFLOP/sec (76% theoretical peak)

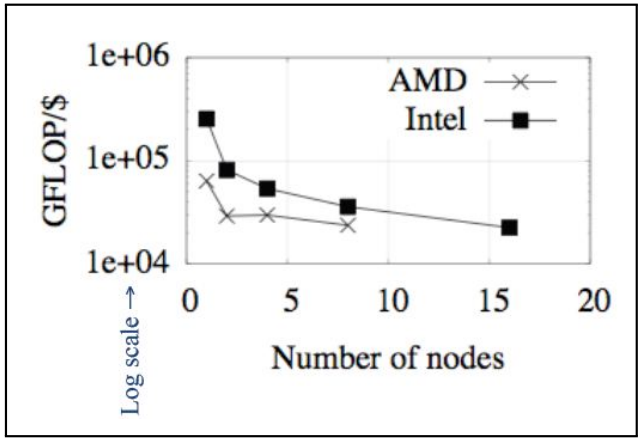
Eficiencia como Função do Ótimo Teórico GFLOPSs - Napper - 2008

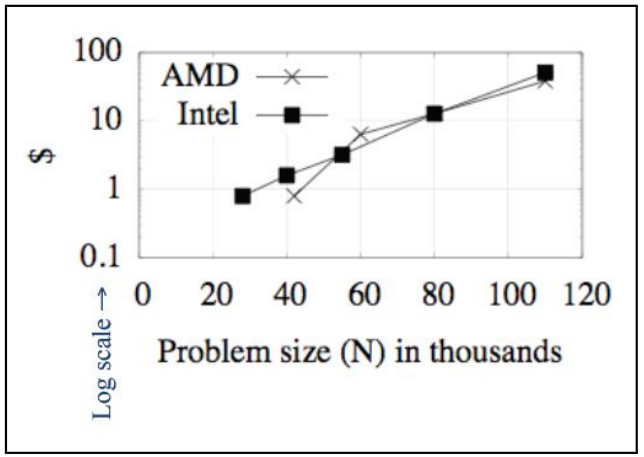
- Considerações sobre a Rede:
  - Parece ser uma interconexão de 1GbE
  - Todas as instancias estão reservadas no mesmo datacenter
  - Conexão WAN parece ser limitada a 20 Mb/s(max.) throughput
  - Vazão da Memória parece Estável / I/O pode ter um grande impacto
  - TCP/IP sob Gigabit ethernet pode apresentar problemas: usou auto-tuning do Linux 2.6 em termos de buffers, e desabilitou TCP slow-start (10% dif.)
- Desempenho não escala devido a limitações de I/O em rede:
  - Comunicação ponto-a-ponto satura em 60MB/s
  - FLOPS fica estagnado em 0.9 GB RAM por core para instancia Intel
  - FLOPS melhora para 3.8 GB RAM por core para instancia AMD
  - Instancia AMD mais balanceada, do ponto de vista de RAM, CPU, rede



Eficiencia como Função de Custo e GFLOPSs - Napper - 2008

- Custos é algo inerente ao modelo cloud:
  - FLOPS sozinho pode ser uma medida insuficiente
  - Desempenho não melhora linearmente (em nossos experimentos), mas os custos aumentam linearmente
  - GFLOP/$ permite comparações entre diferentes provedores de cloud
- GFLOP/$ enfatizam o tempo-para-solução:
  - $/tamanho permite comparação da instâncias com relação a custo
  - Tempo até resolução permite refletir as soluções sob-demanda (cobradas por hora Amazon EC2)
  - Balanceamento: poucos cores com bastante RAM: tempo de solução maior mas mais barato (?), mais cores com menos RAM: tempo de solução mais rapido mas mais caro (?)





Benchmarking Amazon EC2 for High-performance Scientific Computing - Walker - 2008

- Quão efetivo é Cloud comercial para computação cientifica de alto desempenho?
  - Faz uso de macro e micro benchmarks na Amazon EC2 comparado com a performance de um cluster
- Resultados mostram um gap de desempenho
  - Para se tornar viável, a idéia é que o provedores de cloud atualizem seu oferecimento, especialmente na área de provisionamento de rede - interconexão - de alto desempenho para atender esses clientes únicos de HPC

Especificação dos Testes e Comparativo - Walker - 2008

	EC2 High-CPU Cluster	NCSA Cluster
Compute Node	7 GB memory, 4 CPU cores per processor (2.33-GHz Xeon), 8 CPU per node, 64 bits, 1690 GB storage	8 GB memory, 4 CPU cores per processor (2.33-GHz Xeon), 8 CPU per node, 64 bits, 73 GB storage
Network Interconnect	High I/O performance (specific interconnect technology unknown)	Infiniband switch

TABLE 1. HARDWARE SPECIFICATIONS OF EC2 HIGH-CPU INSTANCES AND NCSA ABE CLUSTER.

NAS Parallel Benchmark - Walker - 2008

Program	Description	Size	Memory (Mw)
EP	Embarrassingly parallel Monte Carlo kernel to compute the solution of an integral.	230	18
MG	Multigrid kernel to compute the solution of the 3-D Poisson equation.	2563	59
CG	Kernel to compute the smallest eigenvalue of a symmetric positive definite matrix.	75000	97
FT	Kernel to solve a 3-D partial differential equation using an FFT-based method.	512 × 2562	162
IS	Parallel sort kernel based on bucket sort.	225	114
LU	Computational fluid dynamics application using symmetric successive over-relaxation (SSOR).	1023	122
SP	Computational fluid dynamics application using the Beam-Warming approximate factorization method.	1023	22
BT	Computational fluid dynamics application using an implicit solution method.	1023	96

TABLE 2. NPB CLASS B PROGRAM CHARACTERISTICS

Desempenho usando NPB-OMP Walker - 2008

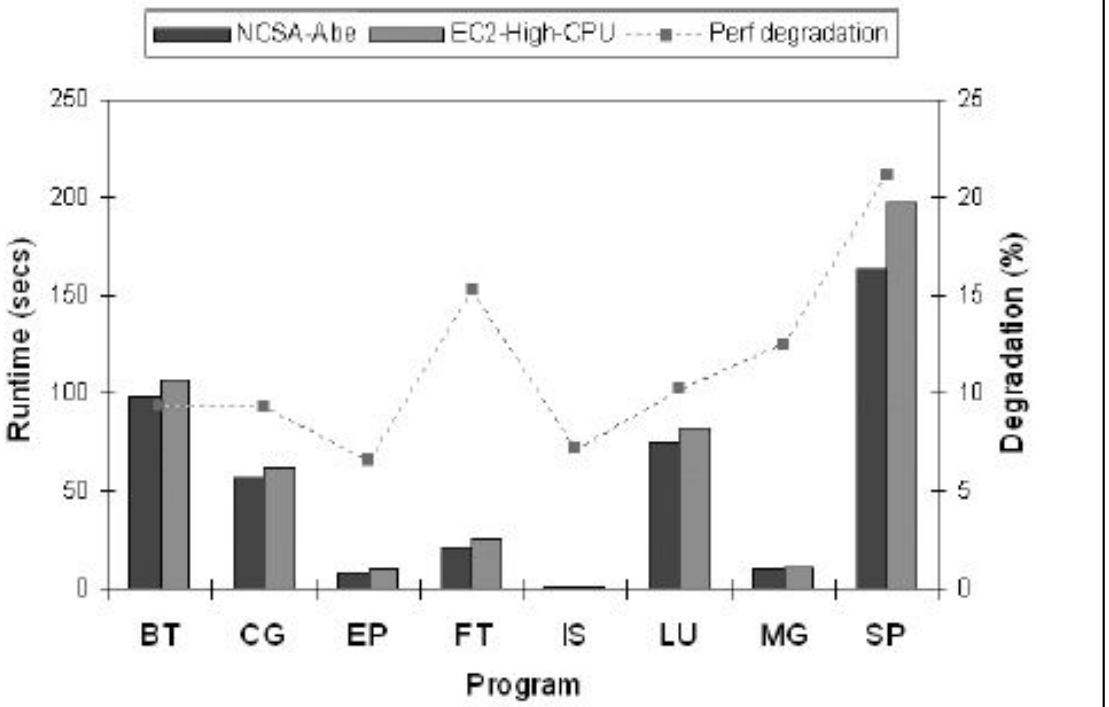


FIGURE 1. NPB-OMP (CLASS B) RUNTIMES ON 8 CPUS ON EC2 AND NCSA CLUSTER COMPUTE NODES. OVERLAID IS THE PERCENTAGE PERFORMANCE DEGRADATION IN THE EC2 RUNS.

- Executando sob uma única instancia high-CPU extra-large
- 8 CPU cores ou 8 threads paralelas no benchmark, opção -openmp -O3
- Degradação de 7% a 21% devido a virtualização nós são equivalente cluster e EC2

Desempenho usando NPB-MPI Walker - 2008

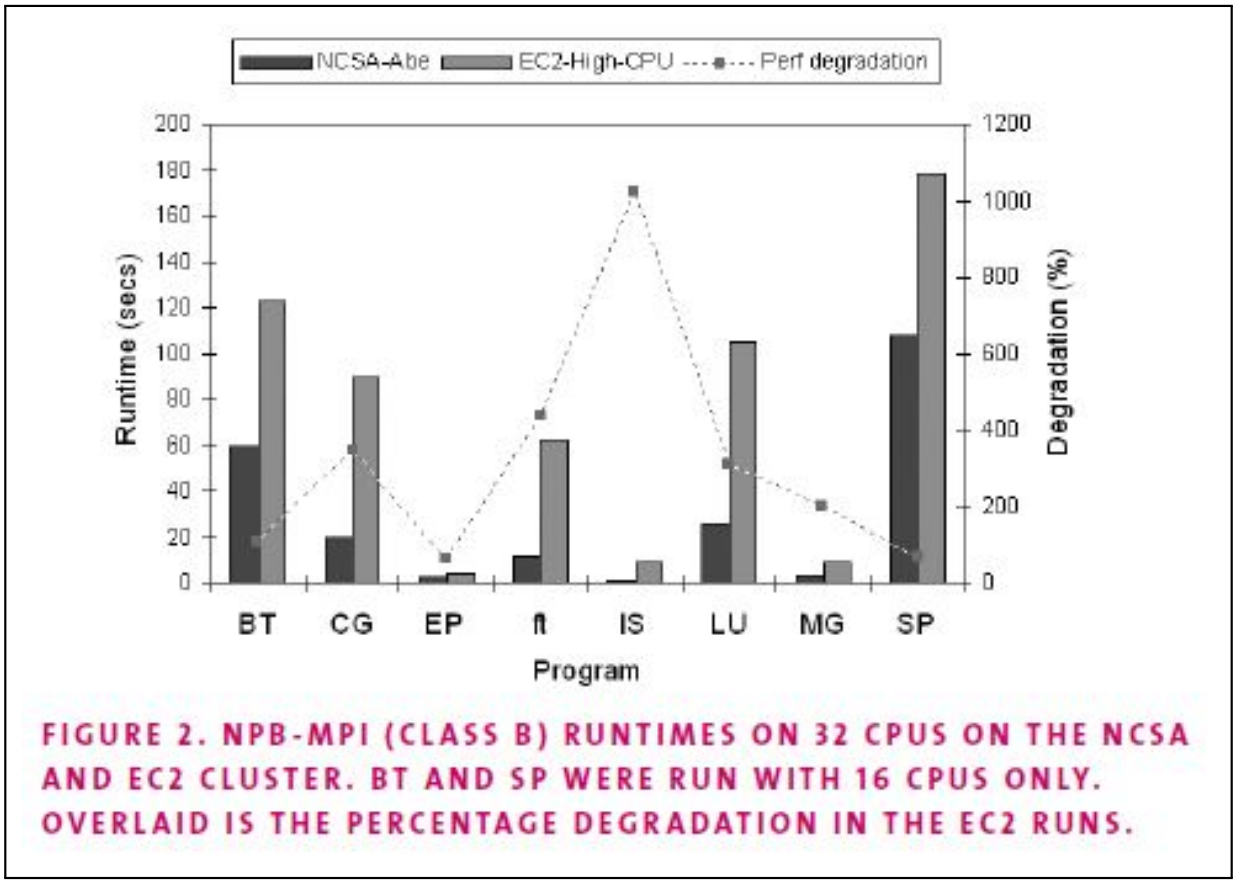


FIGURE 2. NPB-MPI (CLASS B) RUNTIMES ON 32 CPUS ON THE NCSA AND EC2 CLUSTER. BT AND SP WERE RUN WITH 16 CPUS ONLY. OVERLAID IS THE PERCENTAGE DEGRADATION IN THE EC2 RUNS.

- Múltiplos nós de computação - 4 high-CPU extra-large - total 32 CPUs
- Resultados são péssimos em termos de degradação 40% a 1000%
- Mesmo EP (Embarassingly Parallel teve 50% degradação

Hipotése é Infiniband Switch - Walker - 2008

- Execução de testes mpptest benchmark
- Testes de Bisection - o sistema completo é dividido em 2 subsistemas
- Banda e Latencia Agregada é medida com base em mensagem MPI de diferentes tamanhos

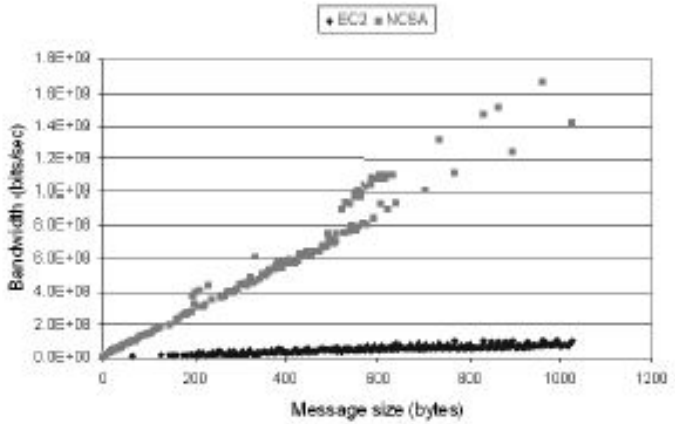


FIGURE 3. MPI BANDWIDTH PERFORMANCE IN THE MPPTEST BENCHMARK ON THE NCSA AND EC2 CLUSTERS

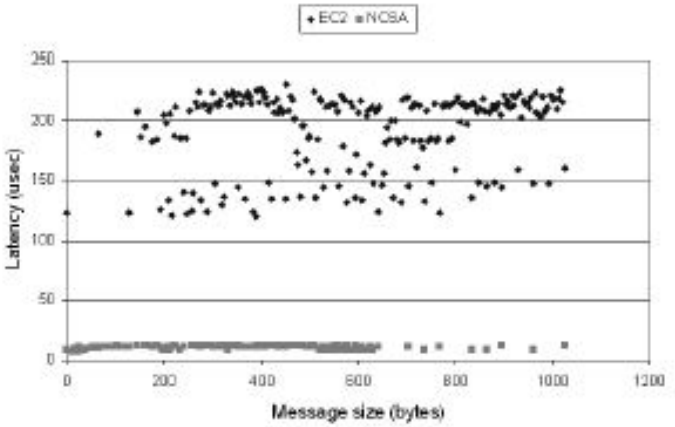
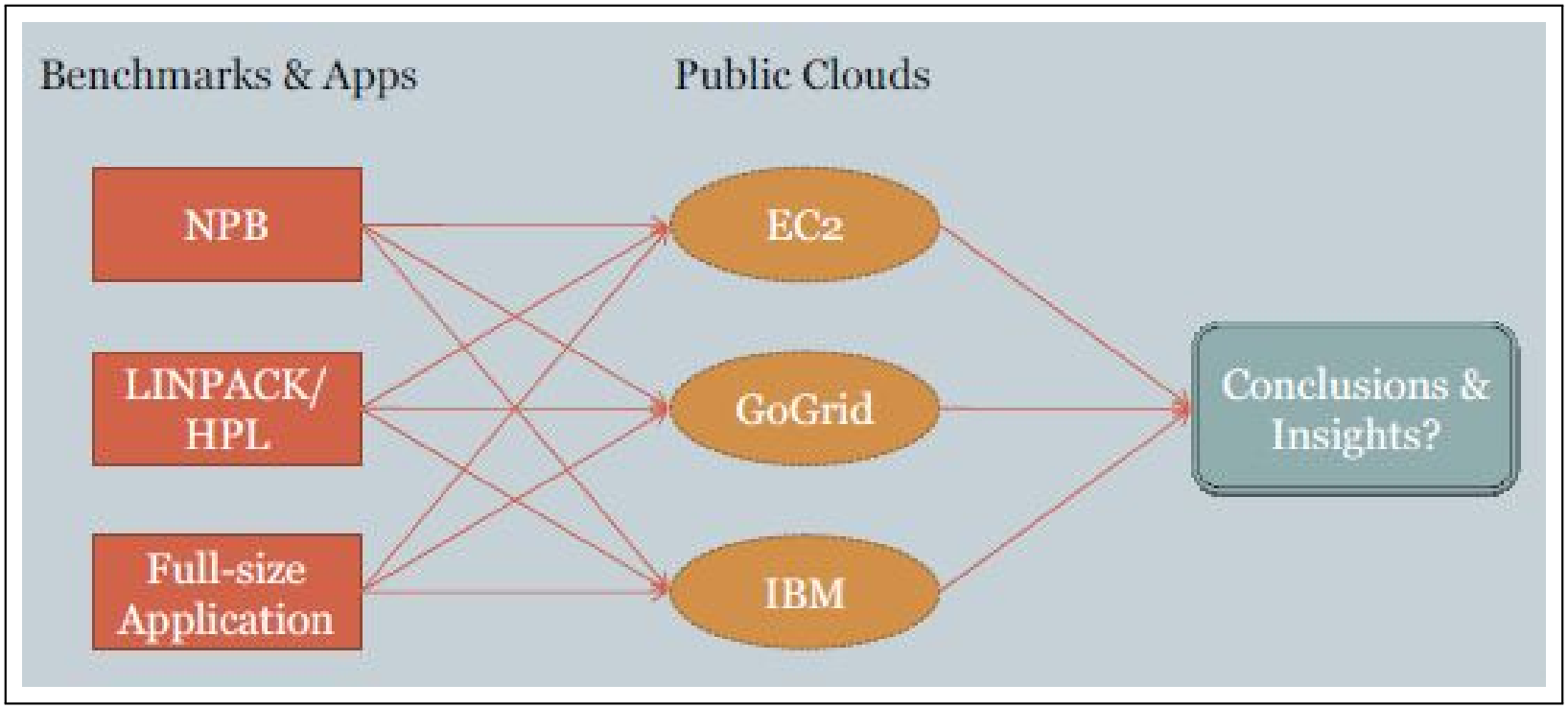


FIGURE 4. MPI LATENCY PERFORMANCE IN THE MPPTEST BENCHMARK ON THE NCSA AND EC2 CLUSTERS

Case Study for Running HPC Applications in Public Clouds - He (Nasa) - 2010

- Idéia é Realizar outros Estudos de Caso Executando HPC em Outras Cloud Públicas
  - Verificar se é técnicamente possível?
  - Se existem clouds mais 'HPC-friendly' que outras
  - Se alguma aplicação é mais 'HPC-friendly'
  - Por que só estudar EC2? Tentar com outras Cloud Públicas
- Resultados mostram que:
  - Paradigma de Programação Importa: MPI, OpenMP, MPI+OpenMP
  - Aplicações Embarrassingly Parallel (EP) tem o melhor desempenho
  - EP ainda assim representam uma ampla variedade de aplicações
  - Quanto a Software de Cloud Comercial versus Open nenhuma diferença foi observada: EC2, GoGrid, IBM

Abordagem de Experimentação - He (Nasa) - 2010

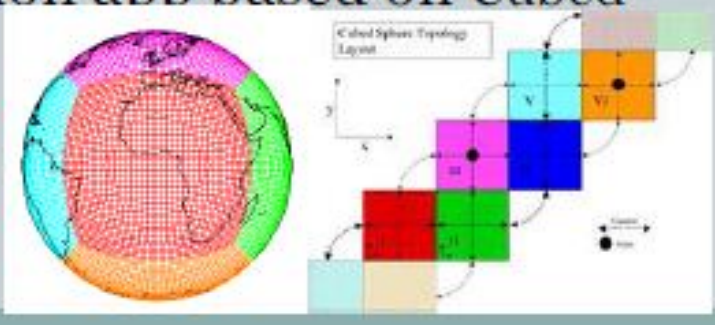


Benchmarks de Teste - He (Nasa) - 2010

- NPB (from NASA)

Benchmark	Name derived from	Description
BT	Block Tridiagonal	Solve a synthetic system of nonlinear PDEs using algorithms block tridiagonal.
CG	Conjugate Gradient	Estimate the largest eigenvalue of a sparse symmetric positive-definite matrix using the inverse iteration with the conjugate gradient method as a subroutine for solving systems of linear equations
EP	Embarrassingly Parallel	Generate independent Gaussian random variates using the Marsaglia polar method
FT	Fast Fourier Transform	Solve a three-dimensional partial differential equation (PDE) using the fast Fourier transform (FFT)
IS	Integer Sort	Sort small integers using the bucket sort
LU	Lower-Upper symmetric Gauss-Seidel	Solve a synthetic system of nonlinear PDEs using LU-SSOR algorithm
MG	MultiGrid	Approximate the solution to a three-dimensional discrete Poisson equation using the V-cycle multigrid method
SP	Scalar Pentadiagonal	Solve a synthetic system of nonlinear PDEs using scalar pentadiagonal algorithm

- LINAPCK/HPL (used by TOP500)
- CSFV (NASA weather prediction app based on Cubed-Sphere Finite-Volume Core)



Especificação de Hardware dos Vários Cloud Providers - He (Nasa) - 2010

- EC2 (Xeon E5345 @2.33GHz)

Server instance	RAM(GB)	Core	Disk(GB)	Price/hr
Small	1.7	1	160	$0.085
Large	7.5	2	850	$0.34
Extra Large	15	4	1690	$0.68
Double Extra Large	34.2	4	850	$1.20
Quad Extra Large	68.4	8	1690	$2.40
High-CPU medium	1.7	2	350	$0.17
High-CPU Extra Large*	7	8	1690	$0.68

- GoGrid(Xeon E5459@3GHz)

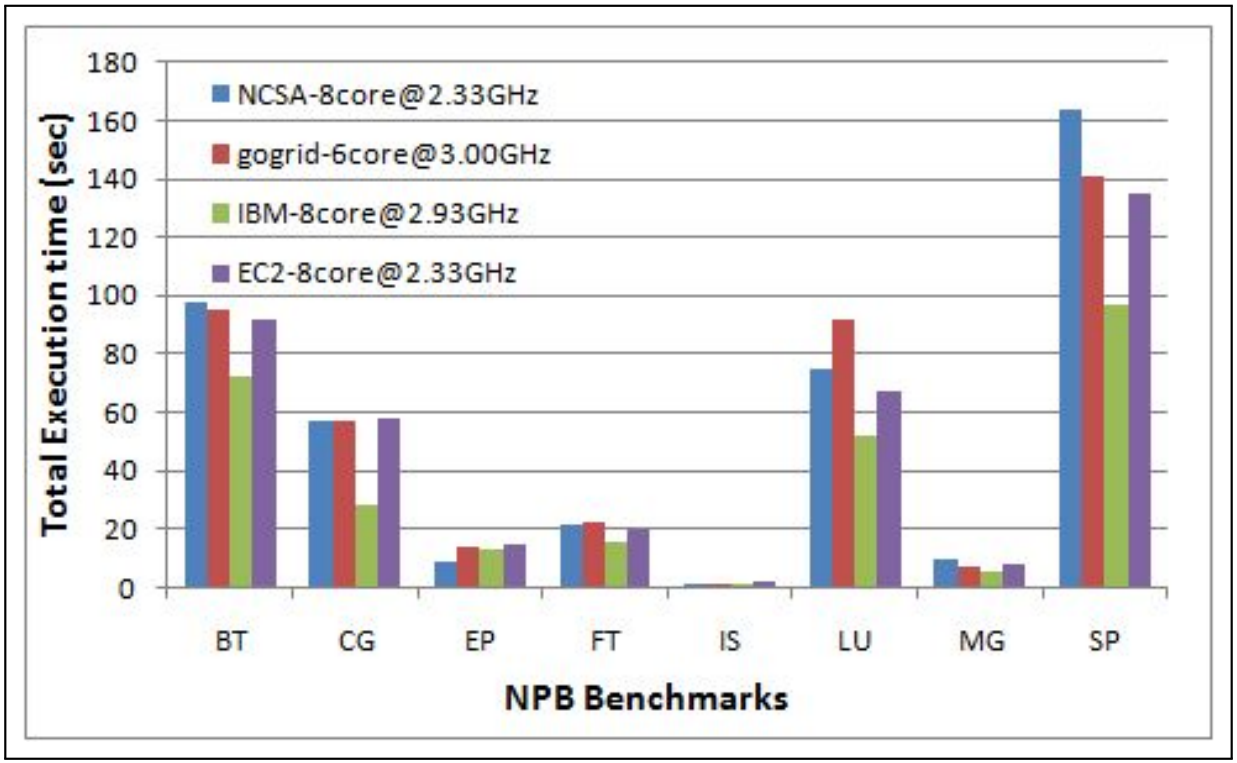
Server Instance	RAM(GB)	Core	Disk(GB)	Price/hr
0.5G	0.5	1	30	$0.095
1G	1	1	60	$0.19
2G	2	1	120	$0.38
4G	4	3	240	$0.76
8G*	8	6	480	$1.52

Now 8 cores

- IBM(Nehalem X5570@2.93GHz)

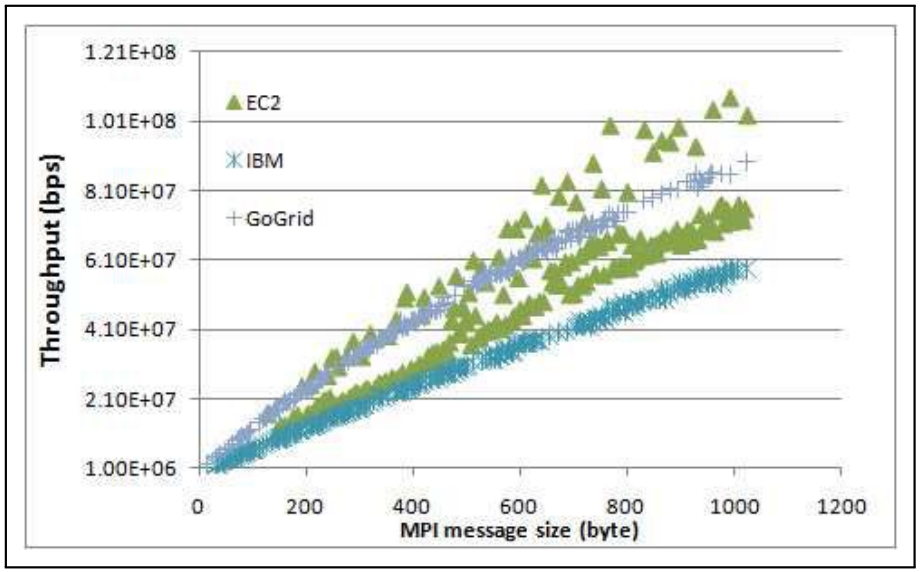
Server Instance	RAM(GB)	Core	Disk(GB)	Price/hr
Small	1	2	8	N/A
Medium	1.8	4	18	N/A
Large*	3.6	8	38	N/A

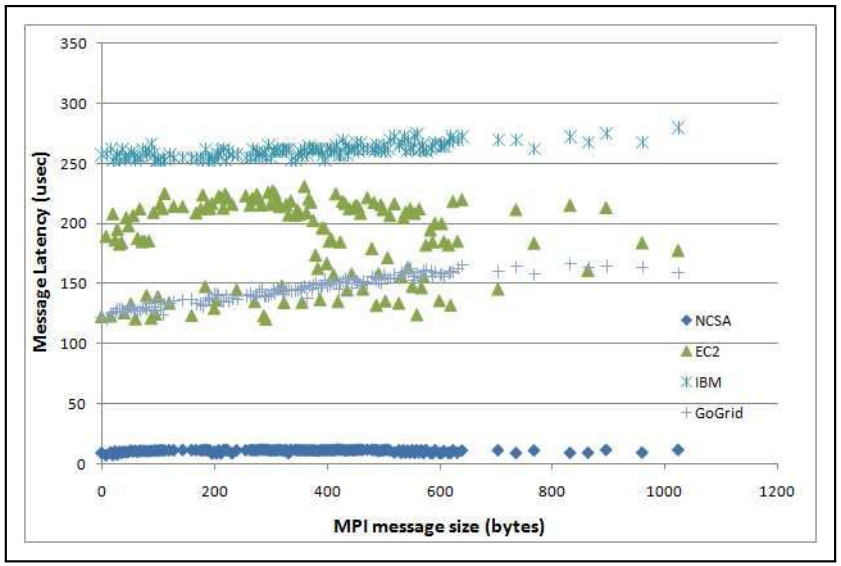
Desempenho usando NPB-OMP - He (Nasa) - 2010



- NPB OpenMP Benchmark (Não executa MPI pela rede)
- Faz a avaliação do overhead de virtualização por nó

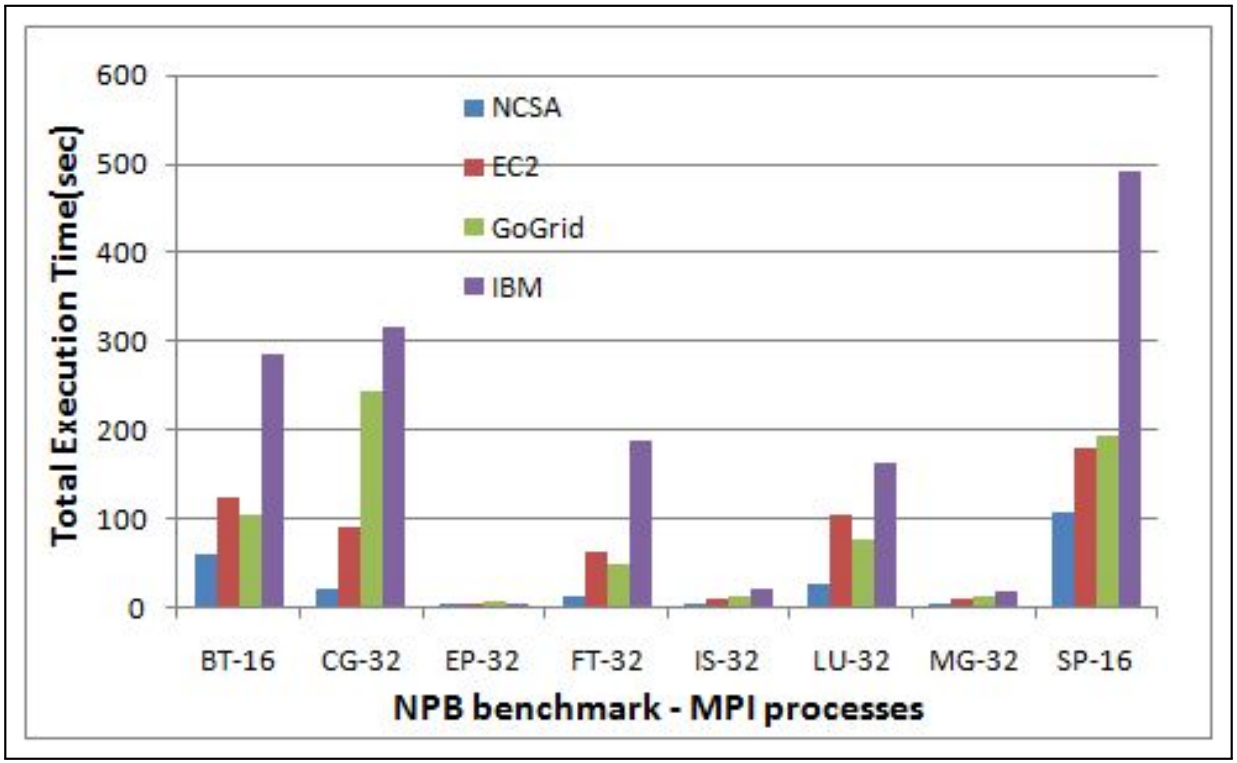
Desempenho usando NPB-MPI - He (Nasa) - 2010





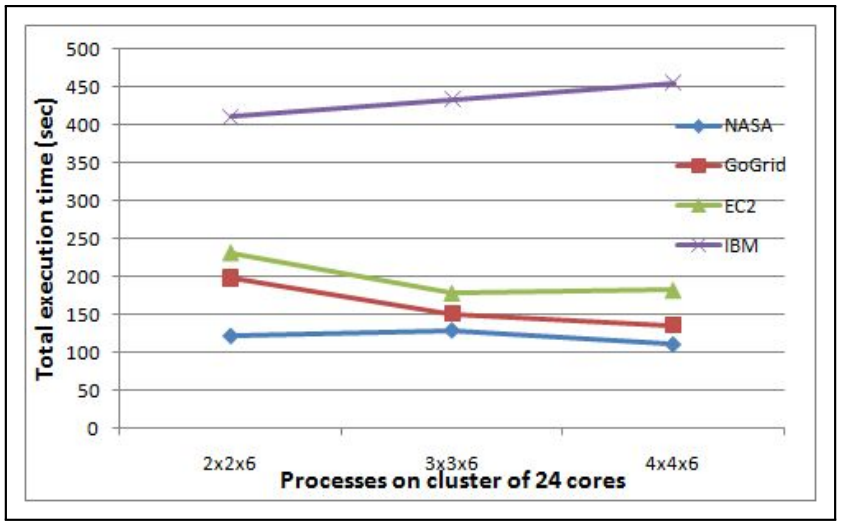
- Com relação a Latencia e Vazão (Mensagens MPI)
- IBM (100Mbps): consistentemente lento
- EC2 (1000Mbps?): inconsistentemente rápido, multi-hop?
- GoGrid (1000Mbps): consistentemente rápido
- Entretanto todos os 3 são ordens de magnitude menores que NCSA (com InfiniBand)

Desempenho usando NPB-MPI - He (Nasa) - 2010



- Re-executou o Benchmark de Walker (2008) nas 3 clouds
- Melhor rede, melhor desempenho (de encontro com pesquisa anterior)
- GoGrid poderia ter sido melhor se tivesse servidores de 8-core

Desempenho usando Aplicação Real CSFV - He (Nasa) - 2010



- Aplicação completa predição climática (com 110K linhas de código Fortran/MPI)
- Cubed-Sphere Finite Volume (CSFV) Dynamic Core - 50 km resolução, simulação 6 horas
- Over-subscription ajuda no desempenho (um pouco) - decomposição em NxNx6 = 24, alternando N
- Cenário de Melhor Caso: somente inferior ao cluster da NASA por 20%, EC2 pior por 50%

E o Futuro de Science Computing em Clouds?

- Amazon EC2 desde 2011 está oferecendo um serviço de High Performance Computing
- Tratam-se de instancias especiais: Cluster Compute e Cluster GPU que funcionam do mesmo jeito que as instancias Amazon EC2 mas tem caracteristicas extras para otimizar o desempenho de aplicações HPC
  - Quando executar um cluster de instancias para prover melhor latencia, é alocado 10Gbps de banda entre as instancias. Clusters com tamanho até 128 instancias é suportado.
  - Instancias Cluster Compute e Cluster GPU podem ser otimizadas pela compilação de aplicações para a arquitetura especifica daquele processador para obter melhor desempenho.
- Instancias Cluster Compute e Cluster GPU estão disponíveis apenas com Linux e em região especifica (US N. Virginia). Precisando de mais que 8 Instancias de Cluster GPU necessário formulário.

Qualidade e Preço de HPC na Amazon

The Cluster Compute instance family currently contains a single instance type, the Cluster Compute Quadruple Extra Large with the following specifications:

- 23 GB of memory
- 33.5 EC2 Compute Units (2 x Intel Xeon X5570, quad-core "Nehalem" architecture)
- 1690 GB of instance storage
- 64-bit platform
- I/O Performance: Very High (10 Gigabit Ethernet)
- API name: cc1.4xlarge

The Cluster GPU instance family currently contains a single instance type, the Cluster GPU Quadruple Extra Large with the following specifications:

- 22 GB of memory
- 33.5 EC2 Compute Units (2 x Intel Xeon X5570, quad-core "Nehalem" architecture)
- 2 x NVIDIA Tesla "Fermi" M2050 GPUs
- 1690 GB of instance storage
- 64-bit platform
- I/O Performance: Very High (10 Gigabit Ethernet)
- API name: cg1.4xlarge

Cluster Compute Instances	
Quadruple Extra Large	$1.60 per hour
Cluster GPU Instances	
Quadruple Extra Large	$2.10 per hour

Amazon HPC

DEMO

Por Amazon AWS usando instancias HPC