HPC in the cloud computing era: challenges, models and tools.

Pascal Bouvry University of Luxembourg, Luxembourg
Mateusz Guzek
Dzmitry Kliazovich
Johnatan E. Pecero
Valentin Plugaru
Sebastien Varrette

Université du Luxembourg

Andrei Tchernykh CICESE Research Center, Mexico

Samee U. Khan North Dakota State University, U.S.A.

Albert Y. Zomaya University of Sydney, Australia

Grand Duchy of Luxembourg

UNIVERSITÉ DU LUXEMBOURG

The Grand Duchy of Luxembourg

- Size:
  - 2,586 km²
- Population:
  - ~ 500,000 inhabitants
  - of which ~ 43 % foreigners
    - of which 50% Portuguese or Italian
  - plus 140.000 commuting
- Capital:
  - Luxembourg
- Official languages:
  - French, German and Luxembourgish

The University

- A new university
  - Created August 2003
  - The one and only in Luxembourg
  - Bologna process right from the start (Bachelor, Master, PhD)

- A multilingual university
  - Three languages (English, German, French)
  - Bilingual and trilingual degrees

- An international university
  - Employees from 20 countries
  - 53% foreign students from 95 countries
  - Over 50 general university agreements for student exchange with universities in Europe, Asia and America as well as 270 ERASMUS agreements for different programmes.
  - Bachelor students have to spend one semester abroad.

uni.lu UNIVERSITÉ DU LUXEMBOURG

UL HPC

2 geographic sites
4 clusters: chaos+gaia, granduc, nyx.
→ 291 nodes, 2944 cores, 27.363 Tflops
→ 1042TB shared storage (raw capa.)
3 system administrators

>5 M USD (Cumul. HW Investment) since 2007

Motivation of the presentation

Ideas in air related to Cloud Computing

- Just pay as you go
- Just drop things in the cloud
- Scale up and down, sky is the limit

Plan

- Cloud
  - History
  - Definition
- Experimental testbed
  - IaaS overhead benchmark
  - Cloud energy model validation
- Greencloud, a cloud infrastructure model and simulator
  - Data center models
  - Network elements
- New CA - DAG model for cloud scheduling
- Conclusion and perspectives

HPC and Cloud

- Historically we had 2 communities:
  - HPC
    - How to benefit from concurrent resources for increasing performance
  - Distributed computing
    - How to remotely access resources, transactional world
- One first attempt to bridge those:
  - Grid computing
    - Public research centers joining forces

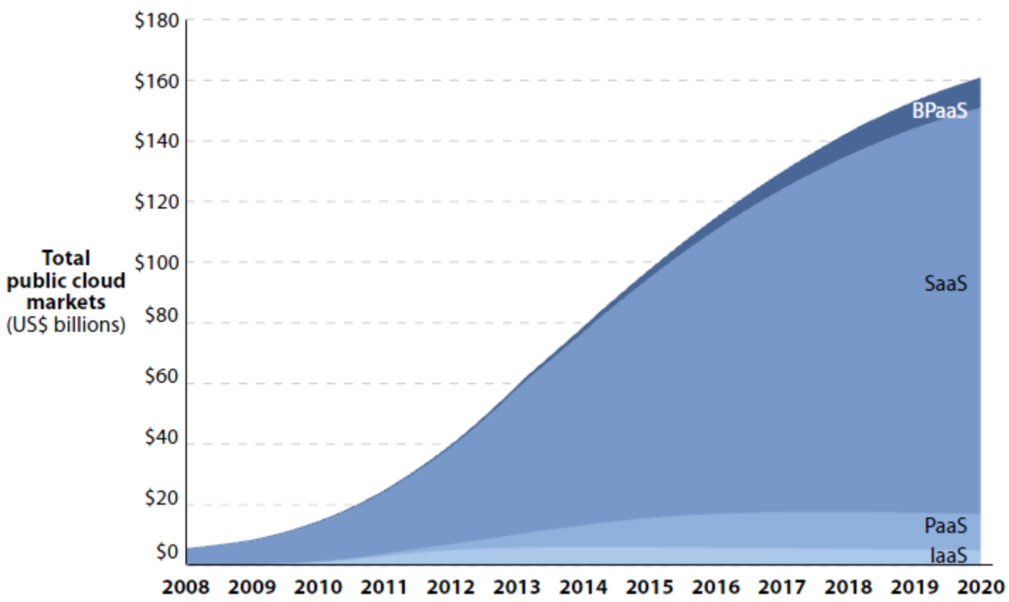
From Grids to Clouds

- Limitations of grids:
  - No real commercial focus (i.e. no clear billing)
  - Complex bundle of various public providers
- Cloud opportunities:
  - Offer coming from the big commercial players
  - Single point of contact providing SLA
  - Virtualization of resources

UNIVERSITÉ DU LUXEMBOURG

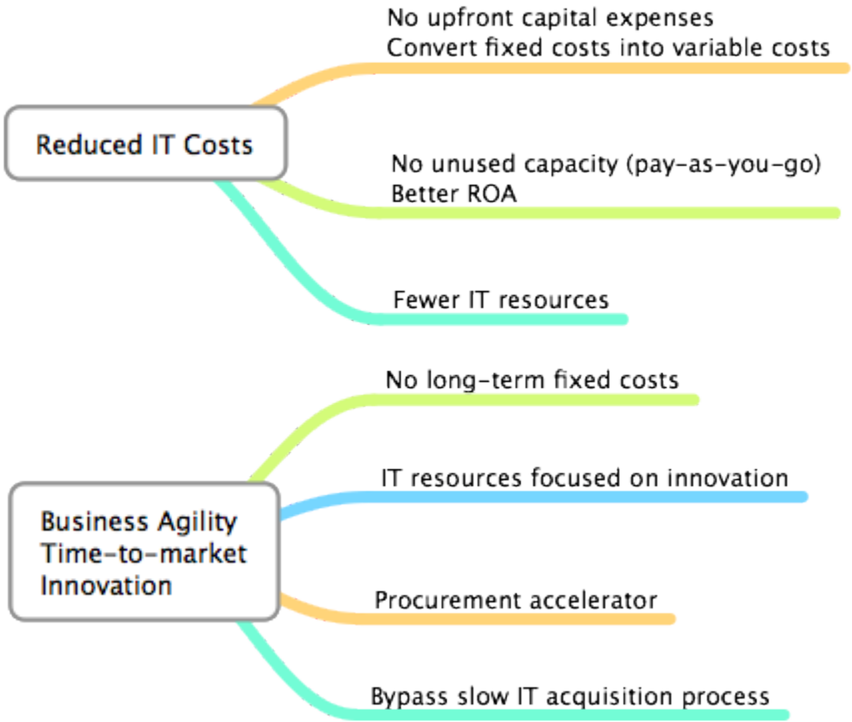
Cloud Computing

- Cloud computing market: $241 billion in 2020
- Main focus is on Software-as-a-Service (SaaS)



Source: Larry Dignan, "Cloud computing market", ZDNet, 2011.

Business Benefits of Cloud Computing



Cloud abstraction layers

Business process
- Business orchestration

Software
- Gmail
- Office365

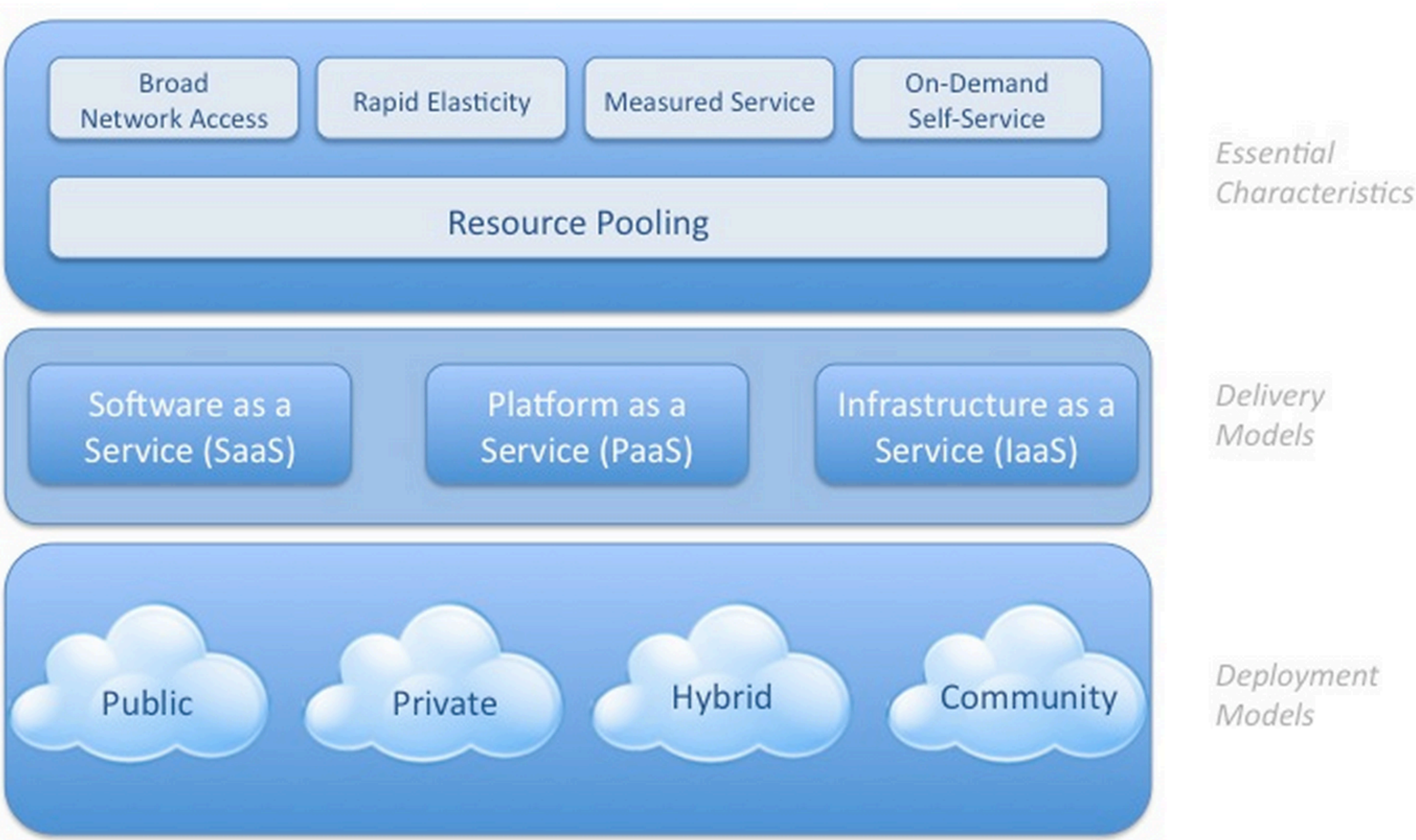
Platform
- Google cloud
- Container based approach

Infrastructure
- Amazon web services
- Virtual machines

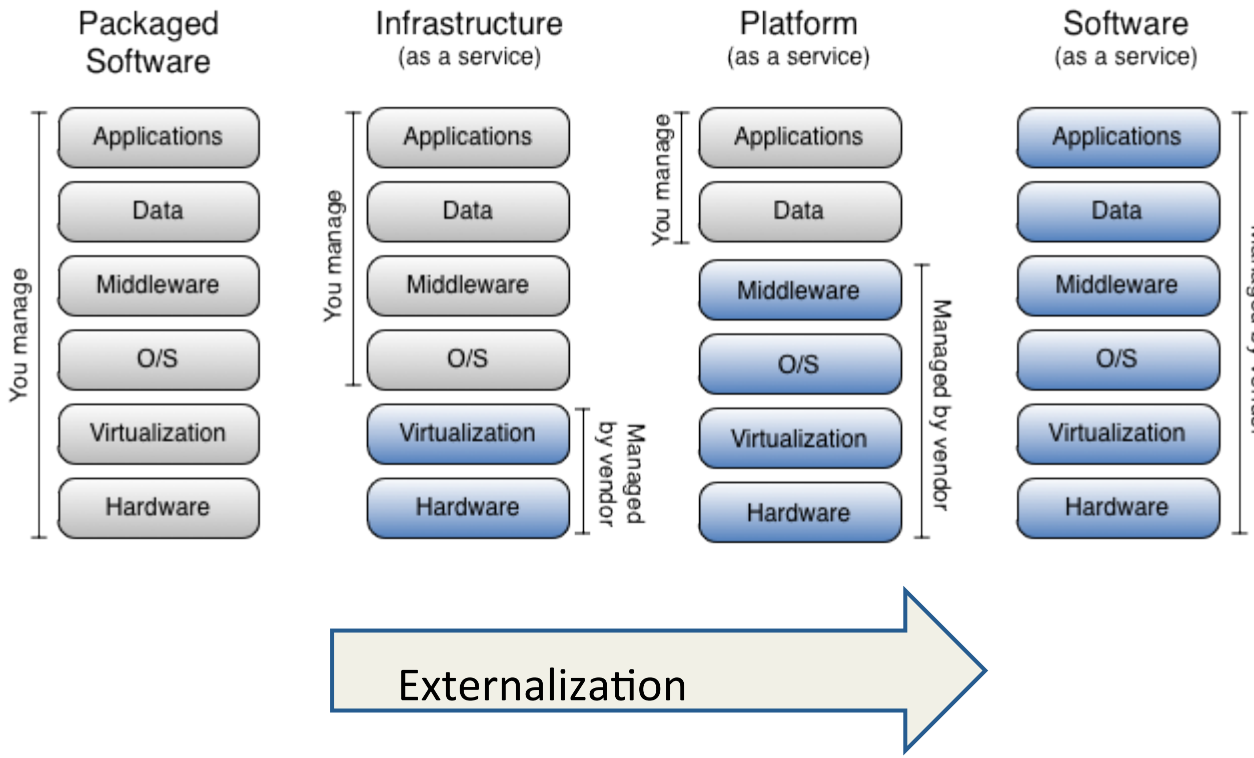
Hardware
- Real hardware
- Hosting

Visual Model Of NIST Working Definition Of Cloud Computing

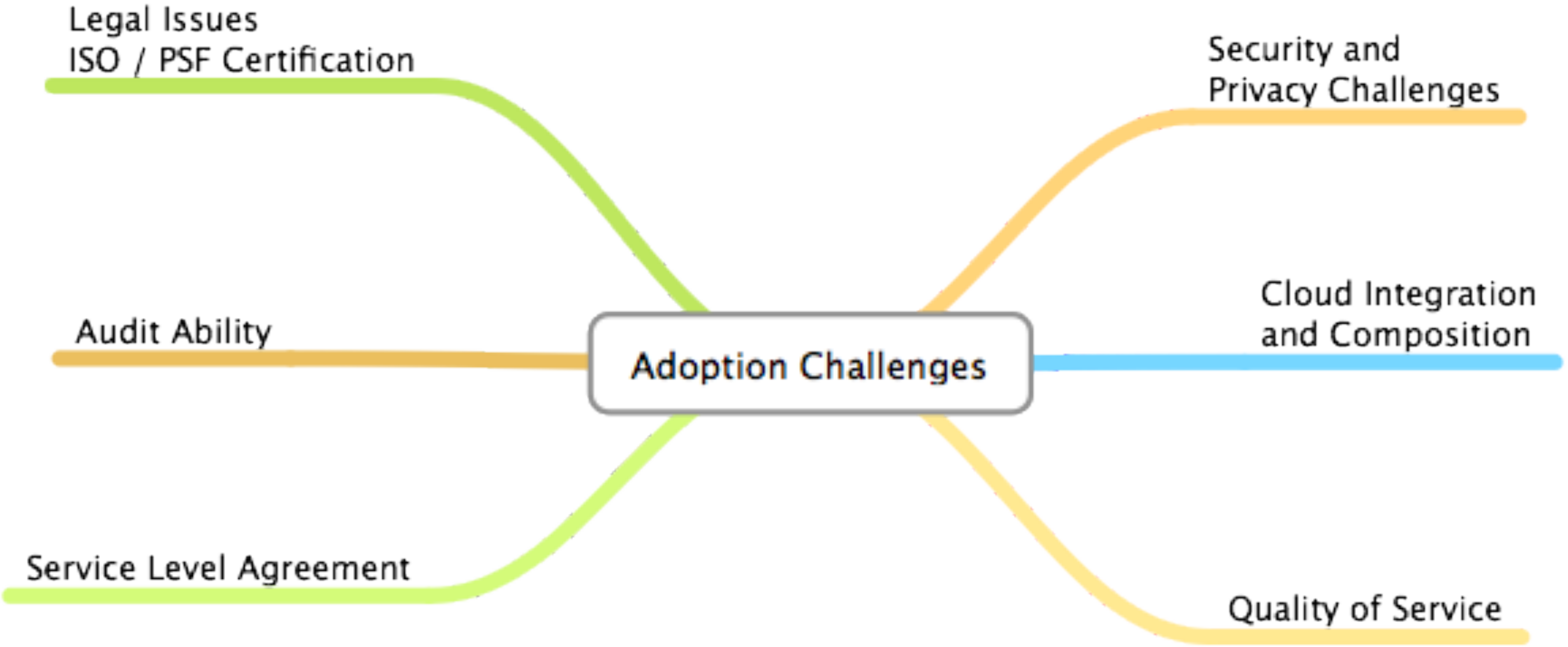
http://www.csrc.nist.gov/groups/SNS/cloud-computing/index.html



Software Stack



Adoption Challenges

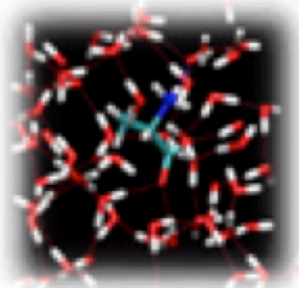


UNIVERSITÉ DU LUXEMBOURG

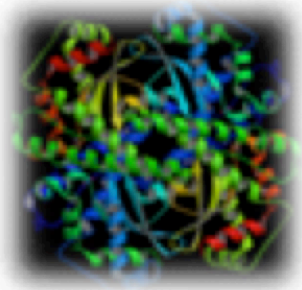
HPC Today

Computational Chemistry
Quantum Mechanics

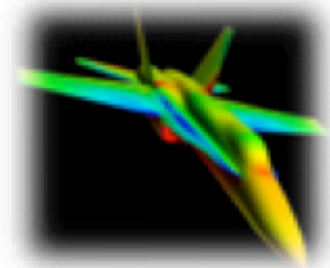
Computational Chemistry
Molecular Dynamics



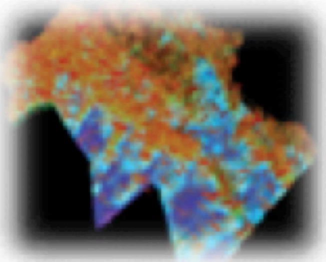
Computational
Biology



Structural Mechanics
Implicit



Reservoir
Simulation



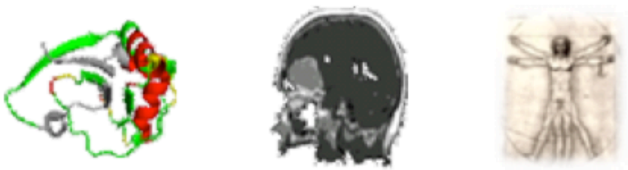
Rendering
Ray Tracing

Climate / Weather
Ocean Simulation

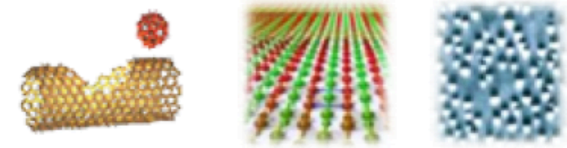
Data Analytics

HPC Tomorrow

Ageing
Medecine
Biology



Materials
Spintronic
Nano-Sciences



- Bridging number crunching and big data!
  - FLOPs but also #Ios, i.e. latency & bandwidth

HPC in the cloud

Horizontal scalability: perfect for replication/ HA (High Availability)

→ best suited for runs with minimal communication and I/O

→ usability for true parallel/distributed HPC runs?

Cloud Data storage

→ Data locality enforced for performance

→ Data outsourcing vs. legal obligation to keep data local

→ Accessibility, security challenges

"Cost effectiveness"

→ chaos+gaia usage: 11,154,125 CPU hours (1273 years) since 2007

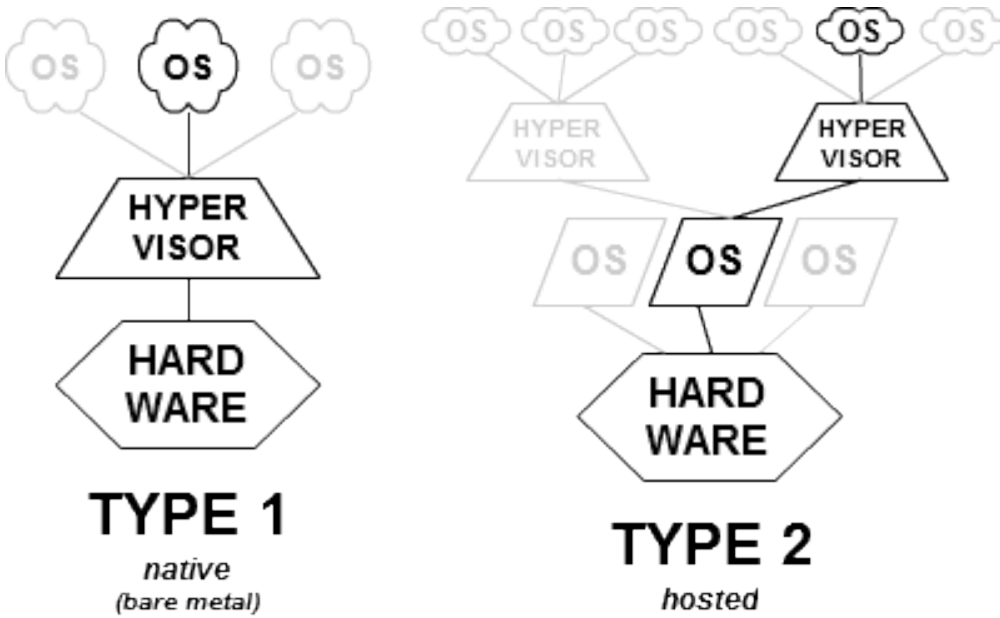
→ 15,06M$ on EC2 cc2.8xlarge vs. 4 Me cumul. HW investment

Virtualization layer impact on performance?

Let's check the virtualization and communication issues

Cloud Middleware Components: Hypervisors

- Hypervisor: core virtualization engine / environment
  - VM running under an hypervisor = *guest* machine
- 2 types of hypervisors
  - native (bare-metal) or hosted



Cloud Middleware Components: Hypervisors

- Native Hypervisors
  - Xen, KVM, ESX[i], Hyper-V
- Hosted hypervisors
  - VMWare Fusion, VirtualBox

Hypervisor:	**Xen 4.0**	**KVM 0.12**	**ESXi 5.1**
Host architecture	x86, x86-64, ARM	x86, x86-64	x86-64
VT-x/AMD-v	Yes	Yes	Yes
Max Guest CPU	128	64	32
Max. Host memory	1TB	-	2TB
Max. Guest memory	1TB	-	1TB
3D-acceleration	Yes (HVM Guests)	No	Yes
License	GPL	GPL/LGPL	Proprietary

Deployment of the same Debian instance on a Grid

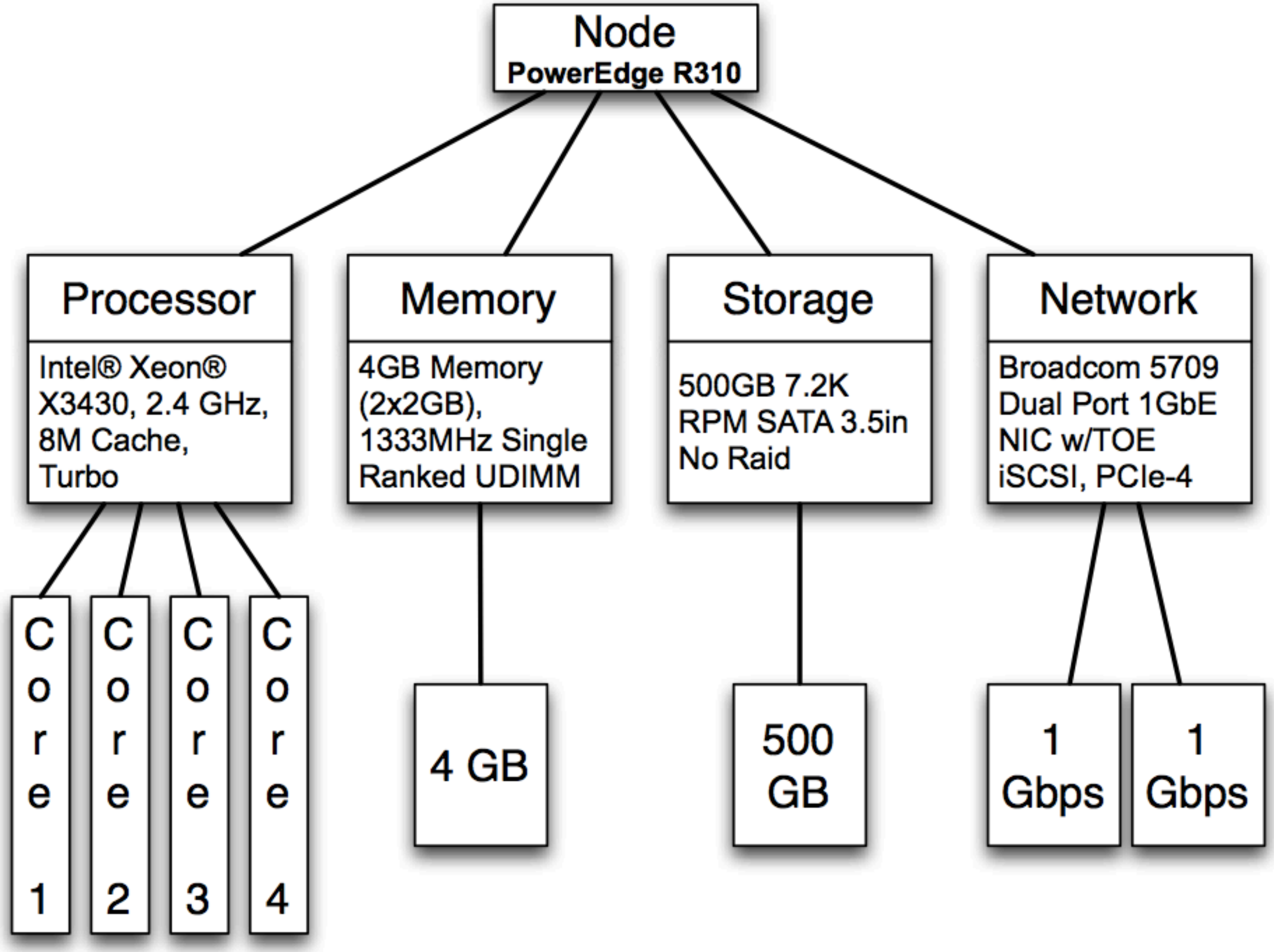
UNIVERSITÉ DU LUXEMBOURG

Benchmark

Selected to represent various use cases of HPC systems:

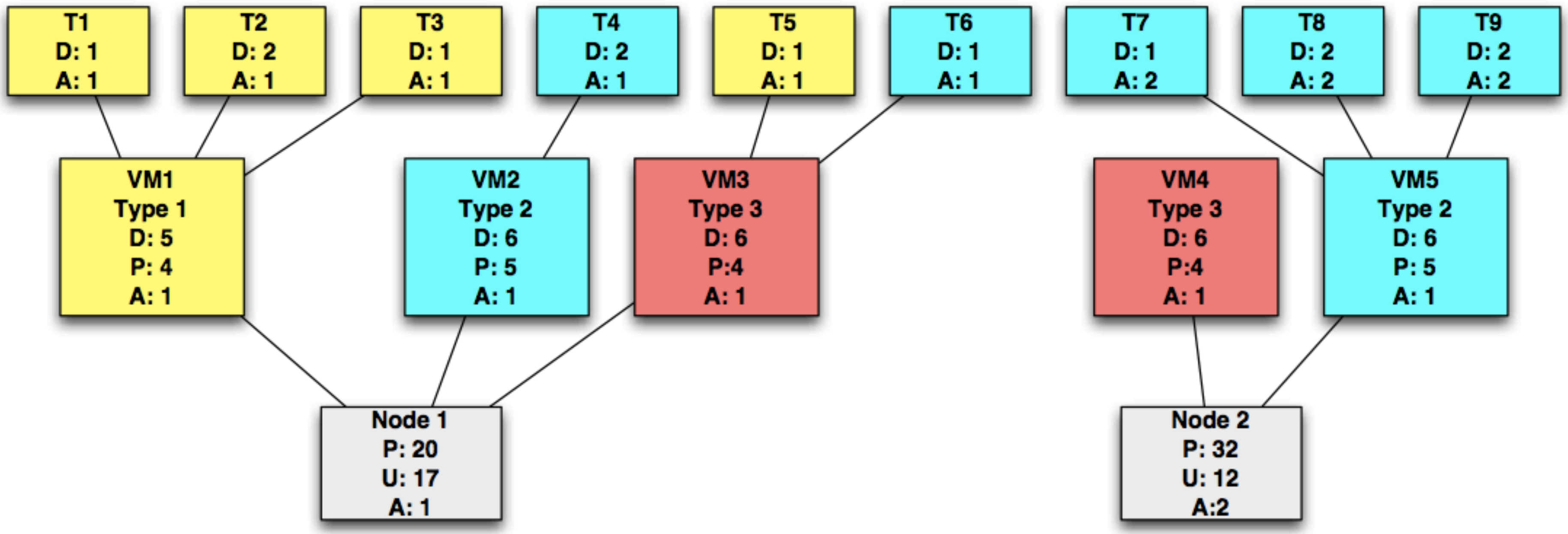
- HPCC : new reference benchmark suit for HPC
  - ↪ includes HPL
  - ↪ 7 tests to stress CPU/disk/RAM/network usage
- Bonnie++ : a **file system** benchmarking suite
- IOZone : cross-platform benchmark of **file operations**
  - ↪ read, write, re-read, re-write, read backwards/strided, mmap. . .

Hardware Model

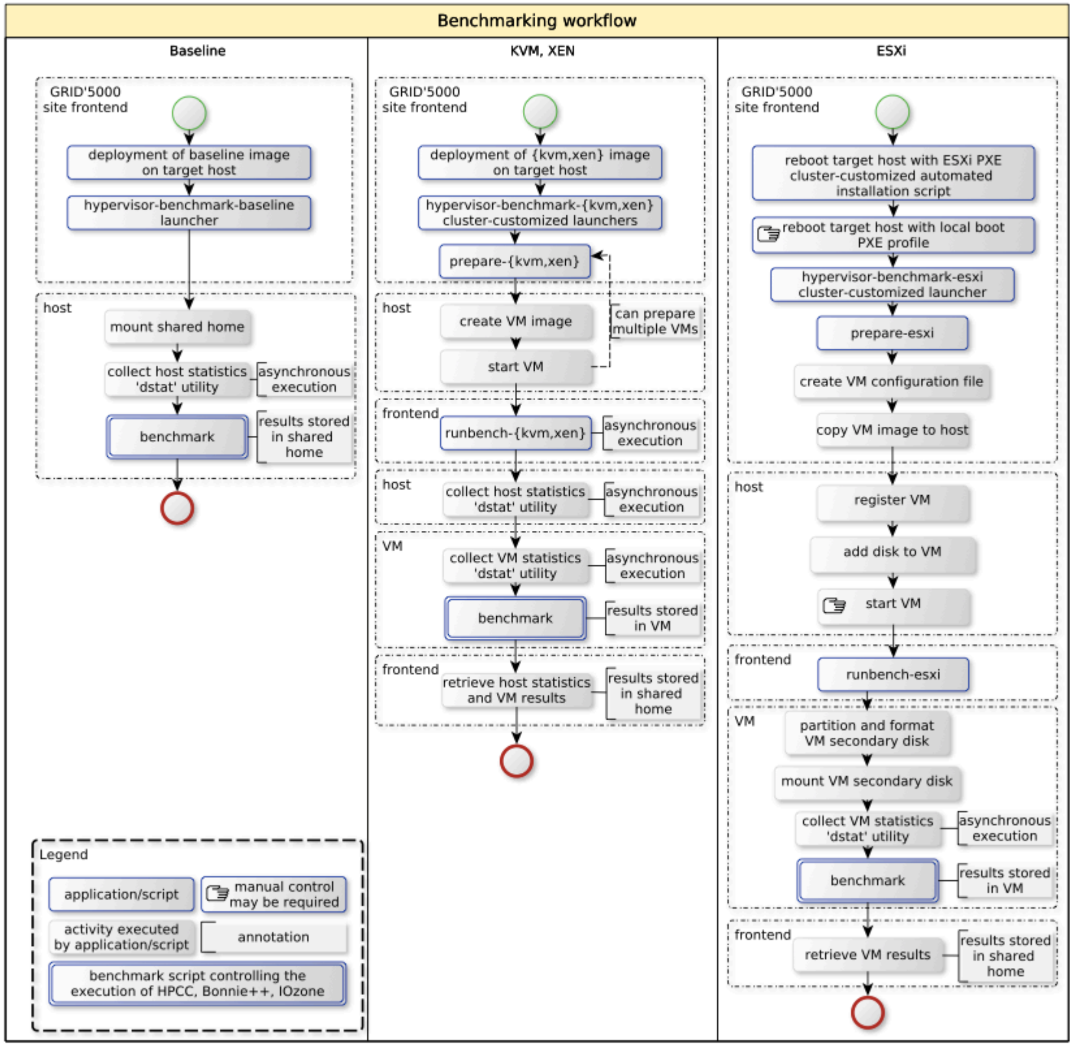


Resource and allocation model

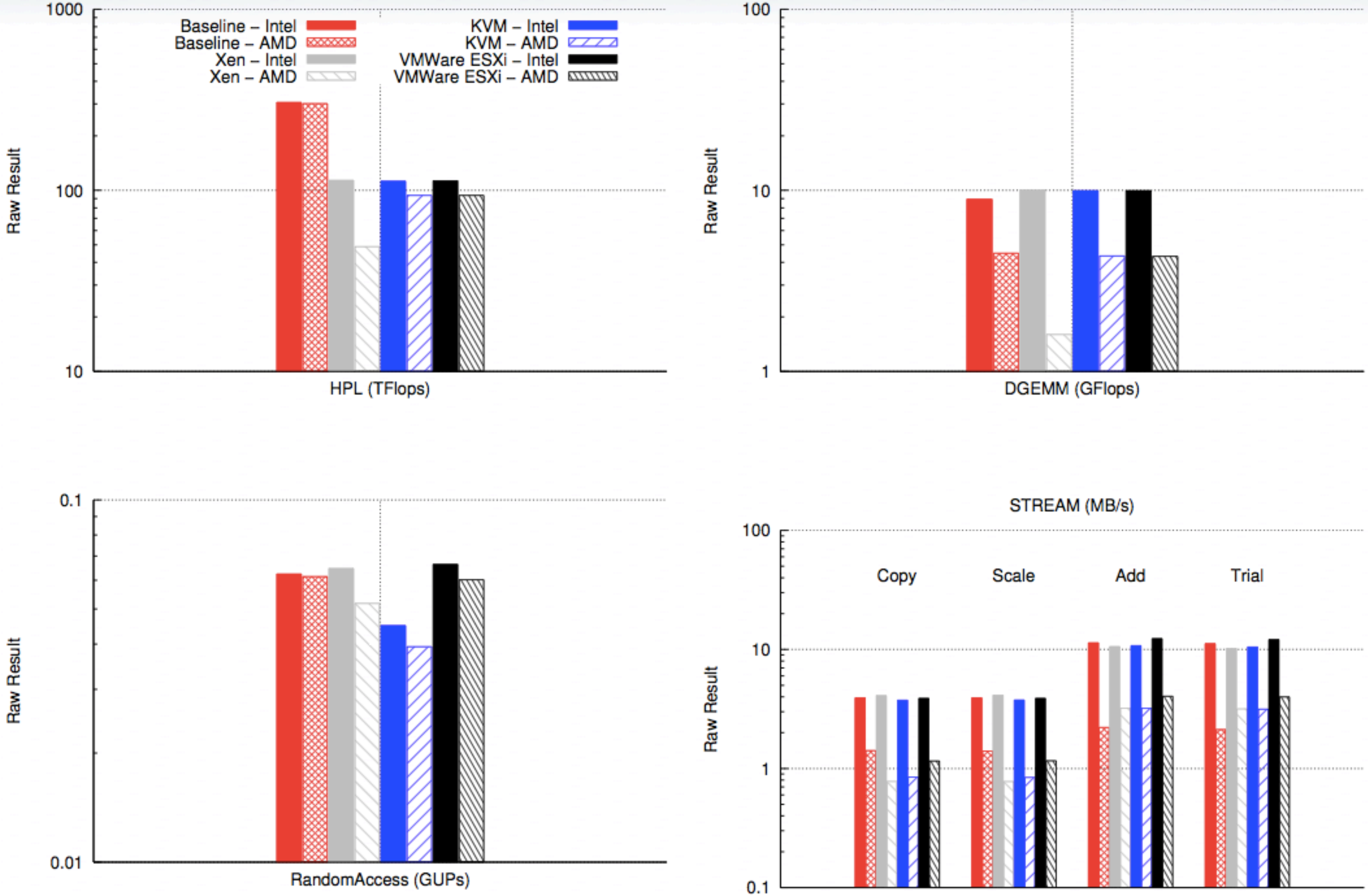
3-tier model: Task, Virtual Machine, Hardware



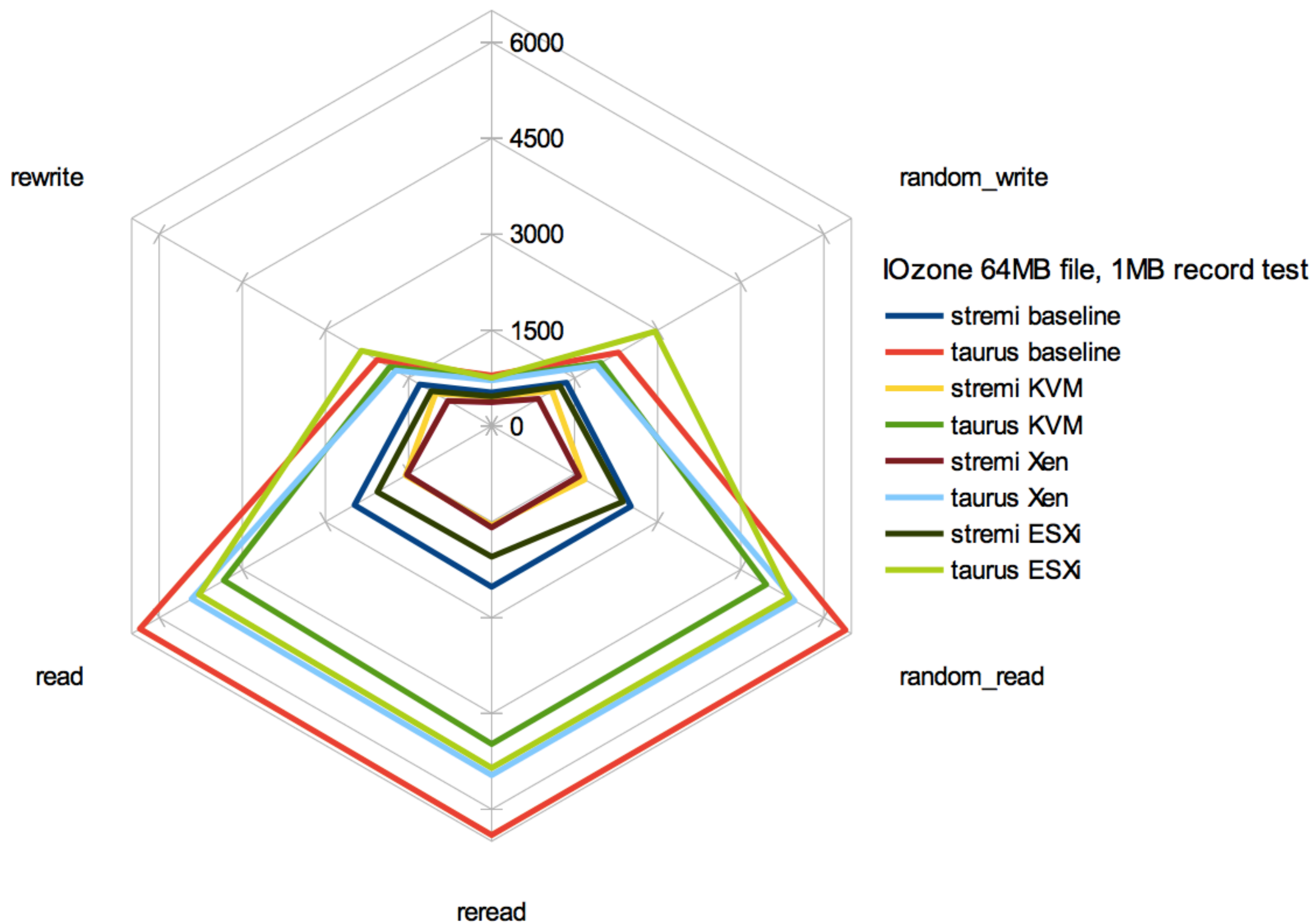
Benchmarking IaaS



Virtualization cost



Virtualization cost



GreenCloud

Simulating Energy-Efficient Clouds

GreenCloud:

A Packet-level Simulator of Energy-aware Cloud Computing Data Centers

http://gforge.uni.lu/greencloud

UNIVERSITÉ DU LUXEMBOURG

Resource Requirements of Cloud Applications

- Computing
- Network Bandwidth
- Communication delays (tolerance)
- Degree of interactivity
- Storage

Resource Requirements of Cloud Applications

- Computing
- Network Bandwidth
- Communication delays (tolerance)
- Degree of interactivity
- Storage

Cloud Computing Applications

Communication resources

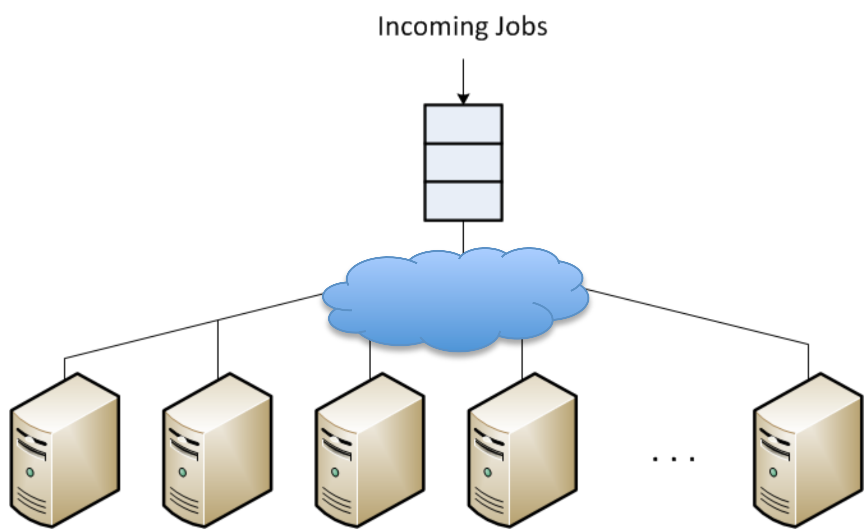
CLASSIFICATION OF CLOUD APPLICATIONS

Cloud application	Resource requirement				
	Computing	Bandwidth	Low communication delay	Degree of interactivity	Storage
Cloud gaming					
Video conferencing					
Online office					
Collaborative editing					
CRM					
Remote desktop					
Cloud Synchronization					
Video streaming					
Cloud storage					
Cloud backup					
Voice conferencing					
Social networking					
HPC					

H: High, M: Medium and L: Low

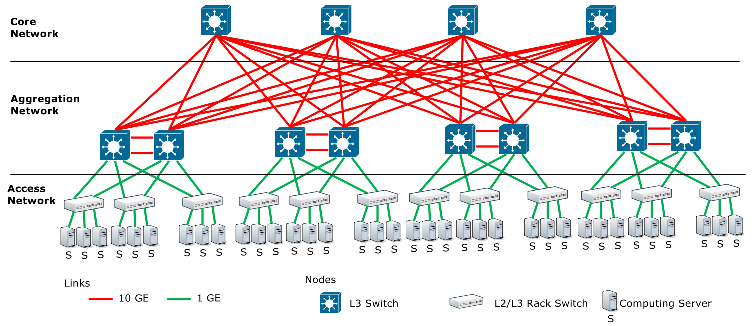
Cloud Computing Applications

- Traditional resource allocation and scheduling
  - Distribute incoming jobs to the pool of servers
  - Communication requirements and networking are not taken into account



GreenCloud: Data Center Architectures

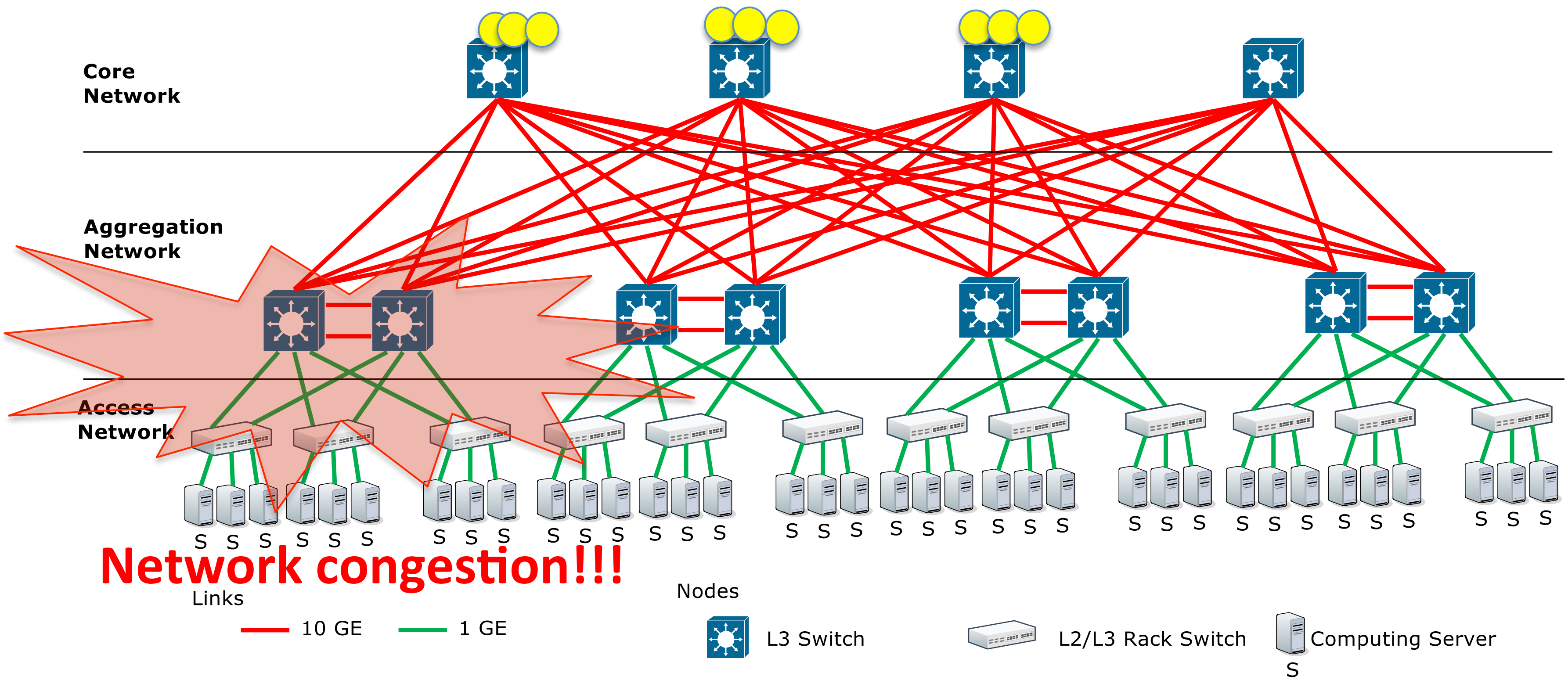
- Supported data center architectures



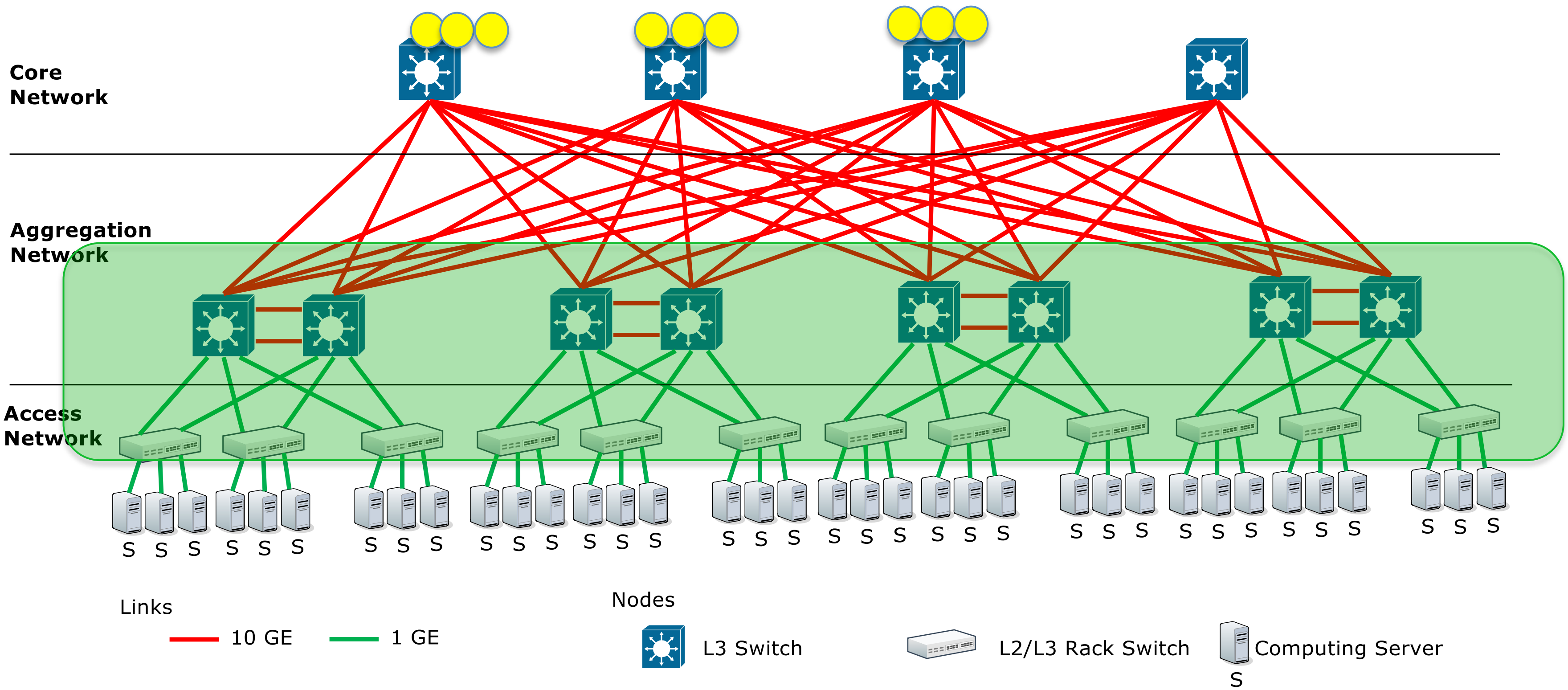
Two/Three-tier data centers

Modular data centers

Scheduling in Data Centers



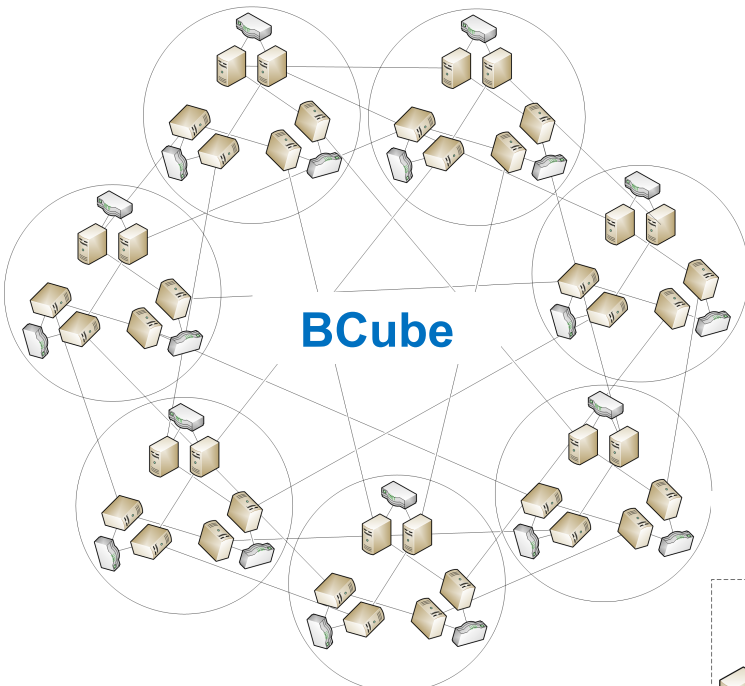
Scheduling in Data Centers

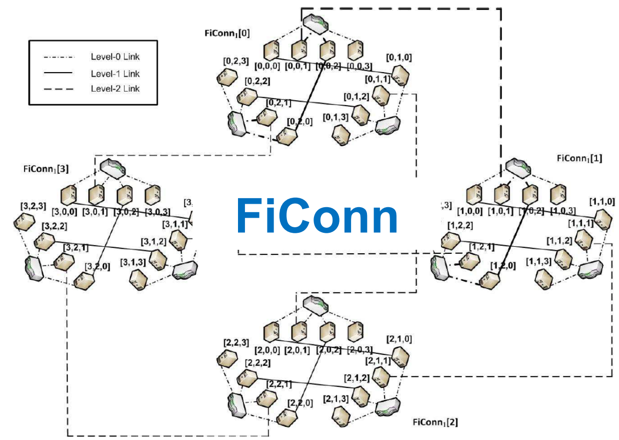


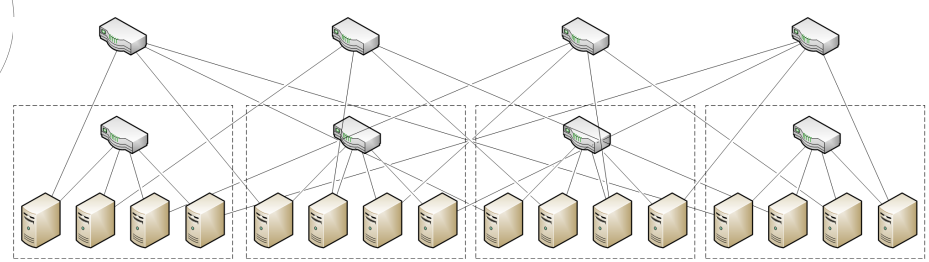
Network is balanced !!!

GreenCloud: Data Center Architectures

- Future data center architectures

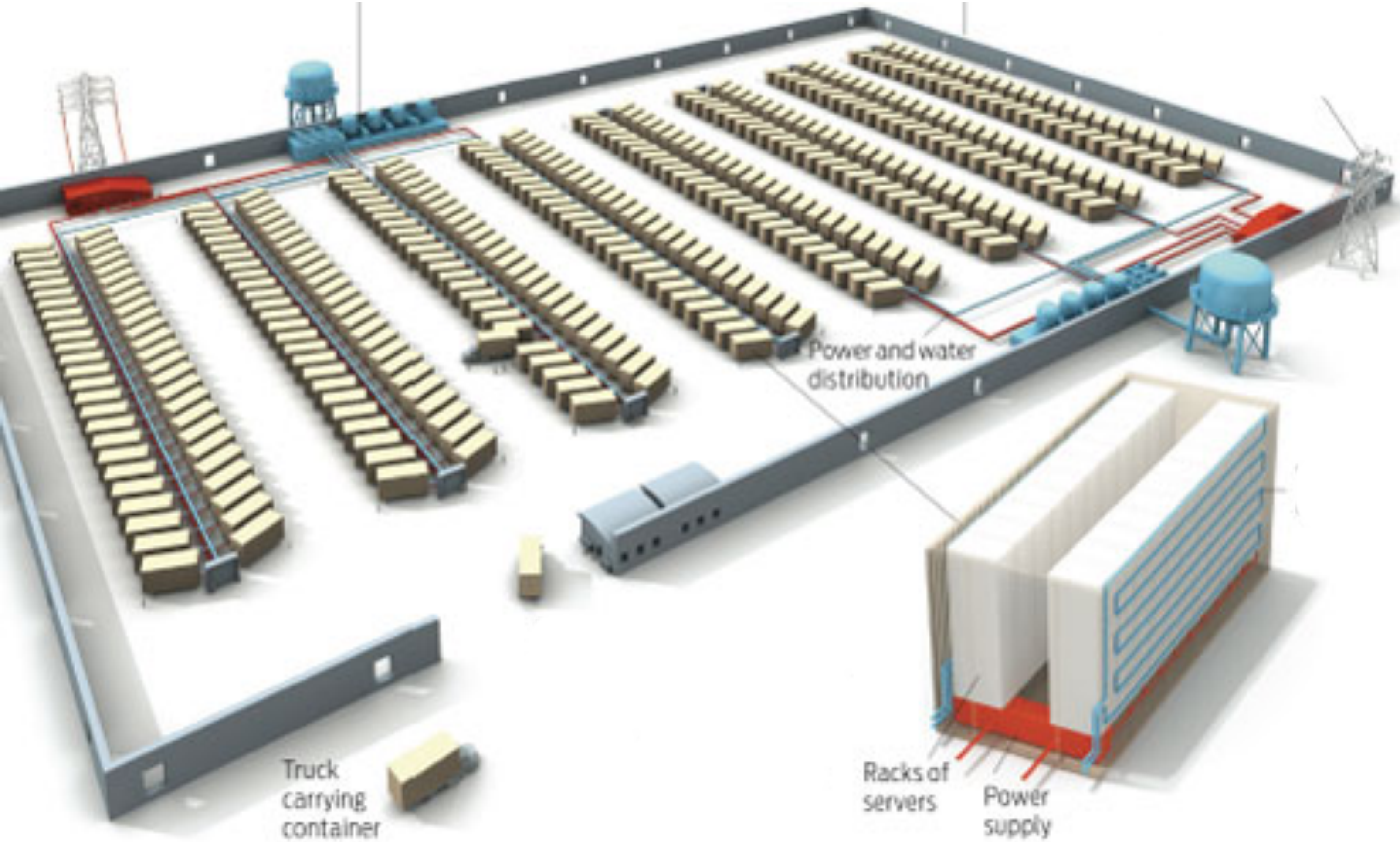






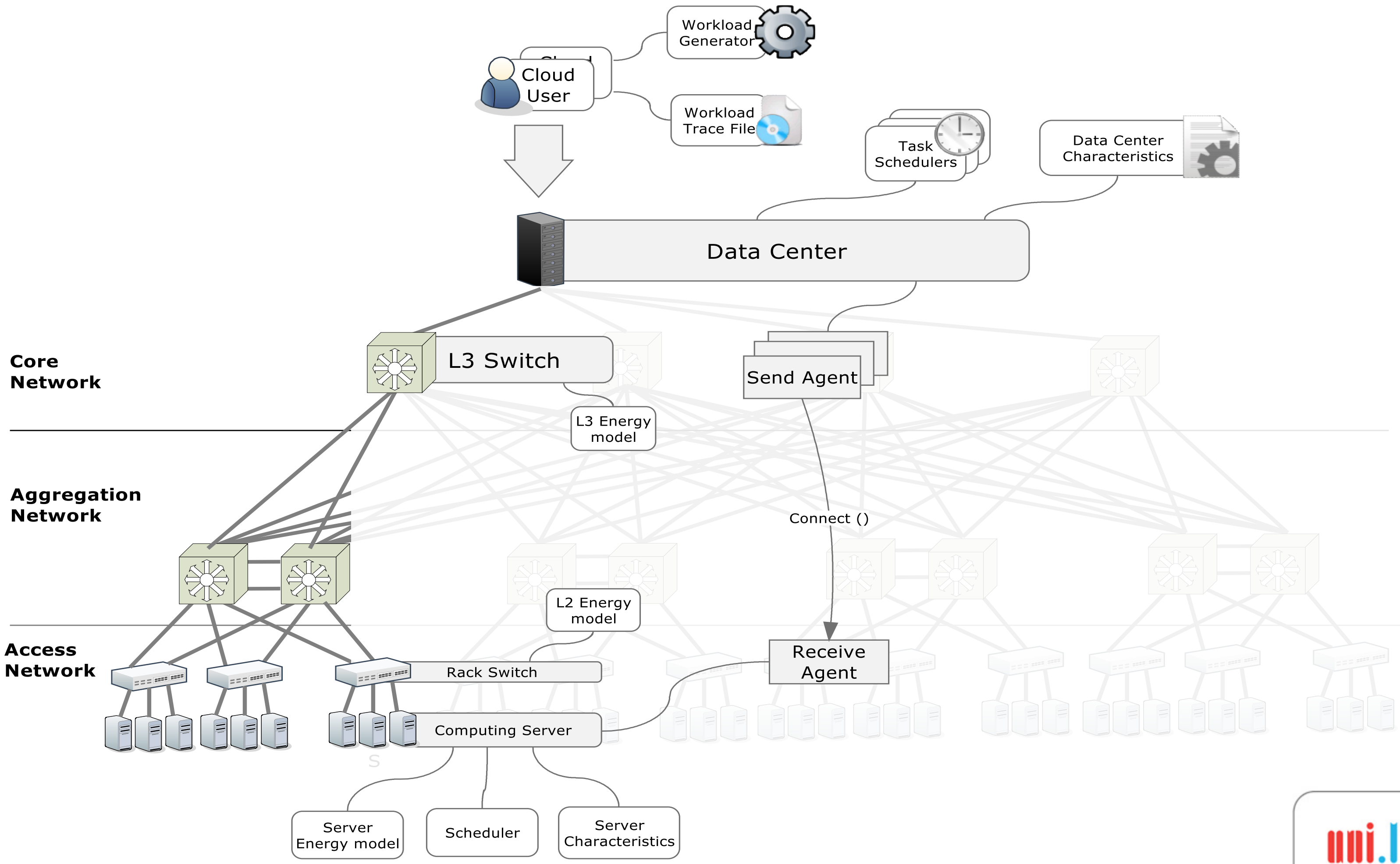
DCell

GreenCloud: Data Center Architectures



Source: Randy Katz, "Tech Titans Boom", *IEEE Spectrum*, June 2009.

GreenCloud Architecture



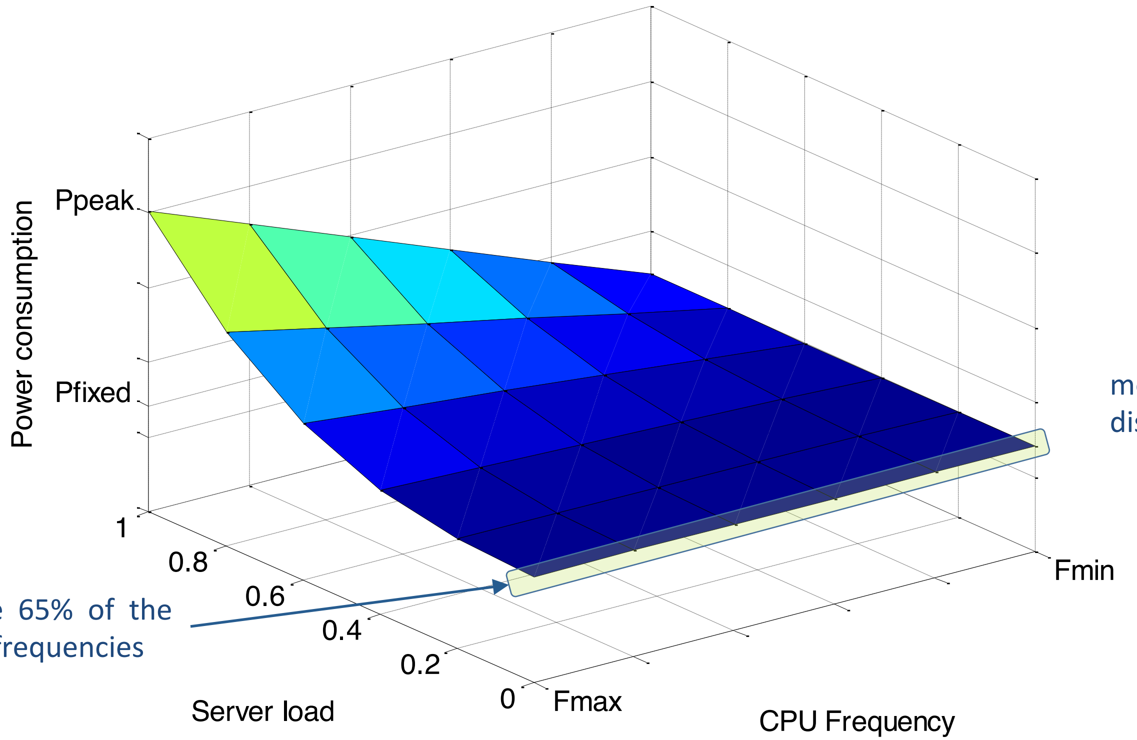
GreenCloud: Simulator Components

- Servers
  - Responsible for task execution
  - Single/multi-core nodes
  - Preset processing limit in MIPS or FLOPS
  - Preset RAM/Disk configuration

- Supported power management modes
  - DVFS: Dynamic Voltage/Frequency Scaling
  - DNS: Dynamic Shutdown (or stand-by)
  - Both: DNS if server is idle, DVFS otherwise

GreenCloud: Simulator Components

- Energy Model for Hosts



$$P = P_{\text{fixed}} + P_f \cdot f^3$$

memory modules, disks, I/O resources

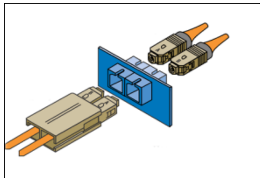
CPU

Idle servers consume 65% of the peak load for all CPU frequencies

GreenCloud: Simulator Components

- Switches
  - Most common Top-of-Rack (ToR) switches typically operate at Layer-2 interconnecting gigabit links in the access network
  - Aggregation and core networks host Layer-3 switches operating at 10 GE (or 100 GE)

- Links
  - Transceivers' power consumption depends on the quality of signal transmission in cables and is proportional to their cost
  - 1 GE links: 0.4W for 100 meter transmissions over twisted pair
  - 10 GE links: 1W for 300 meter transmission over optical fiber



- Supported power management modes
  - DVFS, DNS, or both

GreenCloud: Simulator Components

- Energy model for a network switch

$$P_{switch} = P_{chasis} + n_{linecards} \cdot P_{linecard} + \sum_{i=0}^{R} n_{ports.r} \cdot P_r$$

Chassis
~ 36%

Linecards
~ 53%

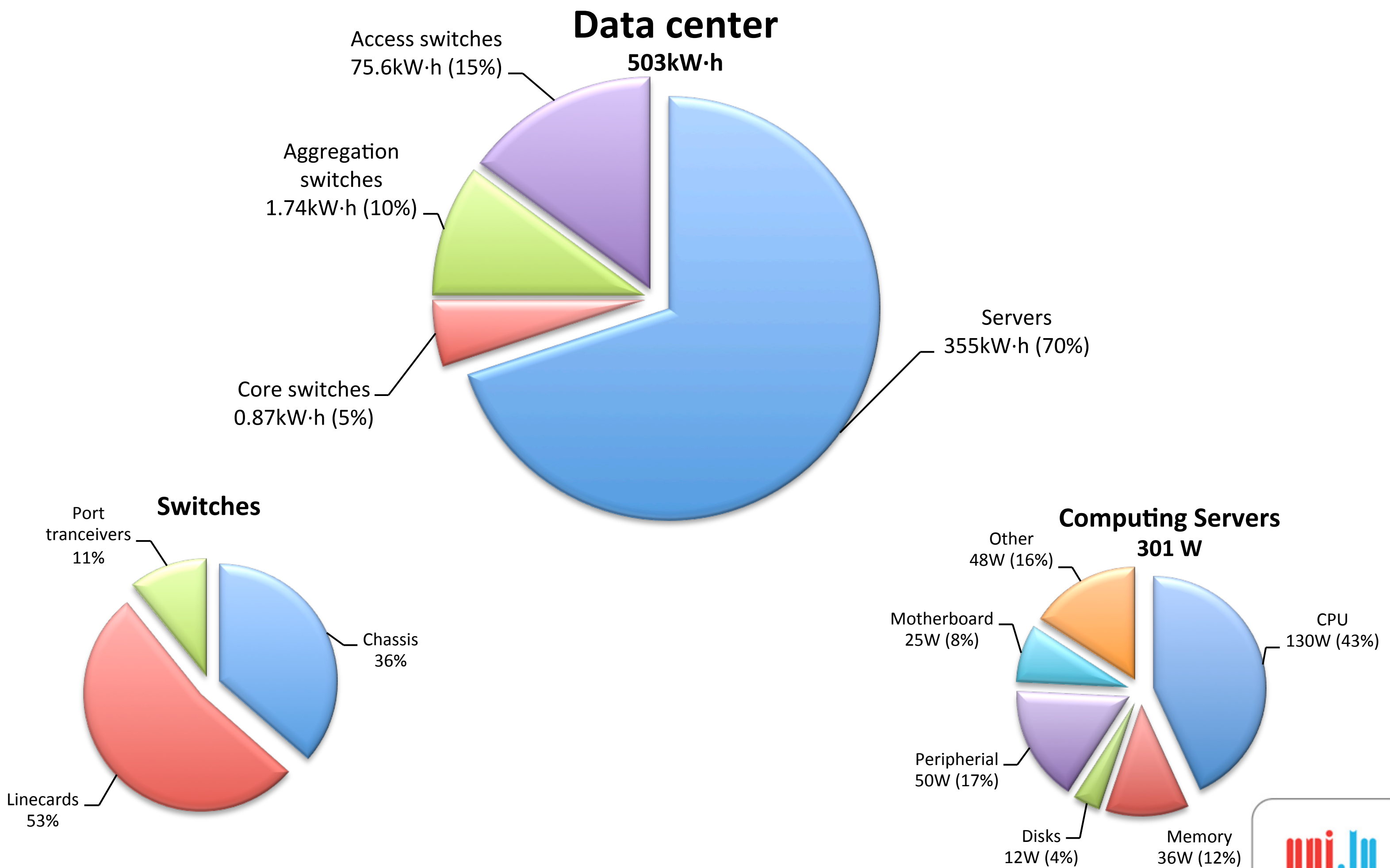
Port transceivers
~ 11%

GreenCloud: Simulator Components

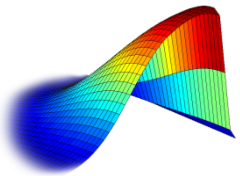
- Workloads
  - Model cloud user applications (social networking, instant messaging, content distribution, etc.)

- Workload properties
  - Computational: MIPS, duration
  - Storage: memory usage
  - Communicational: internal and external communications characteristics
- Generation
  - Trace-driven
  - Using random distribution

GreenCloud: Simulation Results



GreenCloud Innovative Solutions

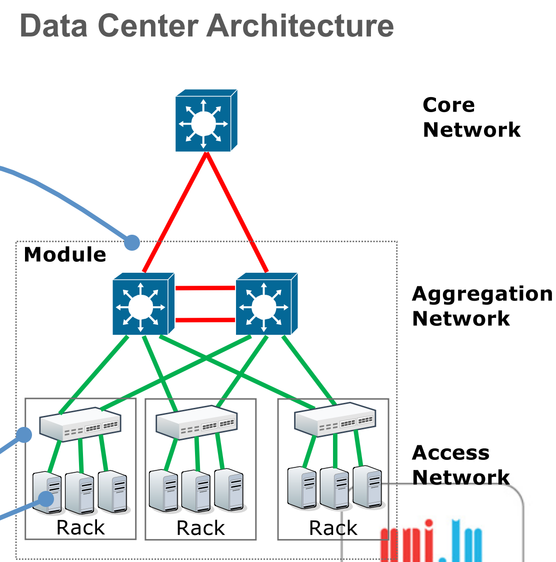


Energy-Efficient Network-Aware Scheduling

– Placing computing jobs to where it will take less energy
– Balance between energy and performance
– IEEE/ACM GreenCom [Best paper award]

DENS is architecture specific

$$M = \alpha \cdot f_s + \beta \cdot f_r + \gamma \cdot f_m$$



GreenCloud Innovative Solutions

e-STAB: Energy-Efficient Scheduling for Cloud Computing Applications with Traffic Load Balancing

#1
- Treat communication and computing demands equally

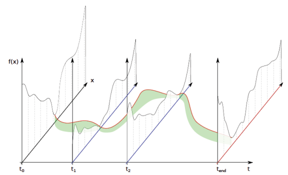
#2
- Optimize energy efficiency and load balancing of network traffic

#3
- Formal model for selection of servers, racks, and network modules

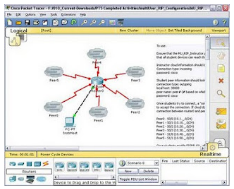
GreenCloud Usage and Benefits

- GreenCloud tools cover complete optimization workflow

1: Client data center analysis



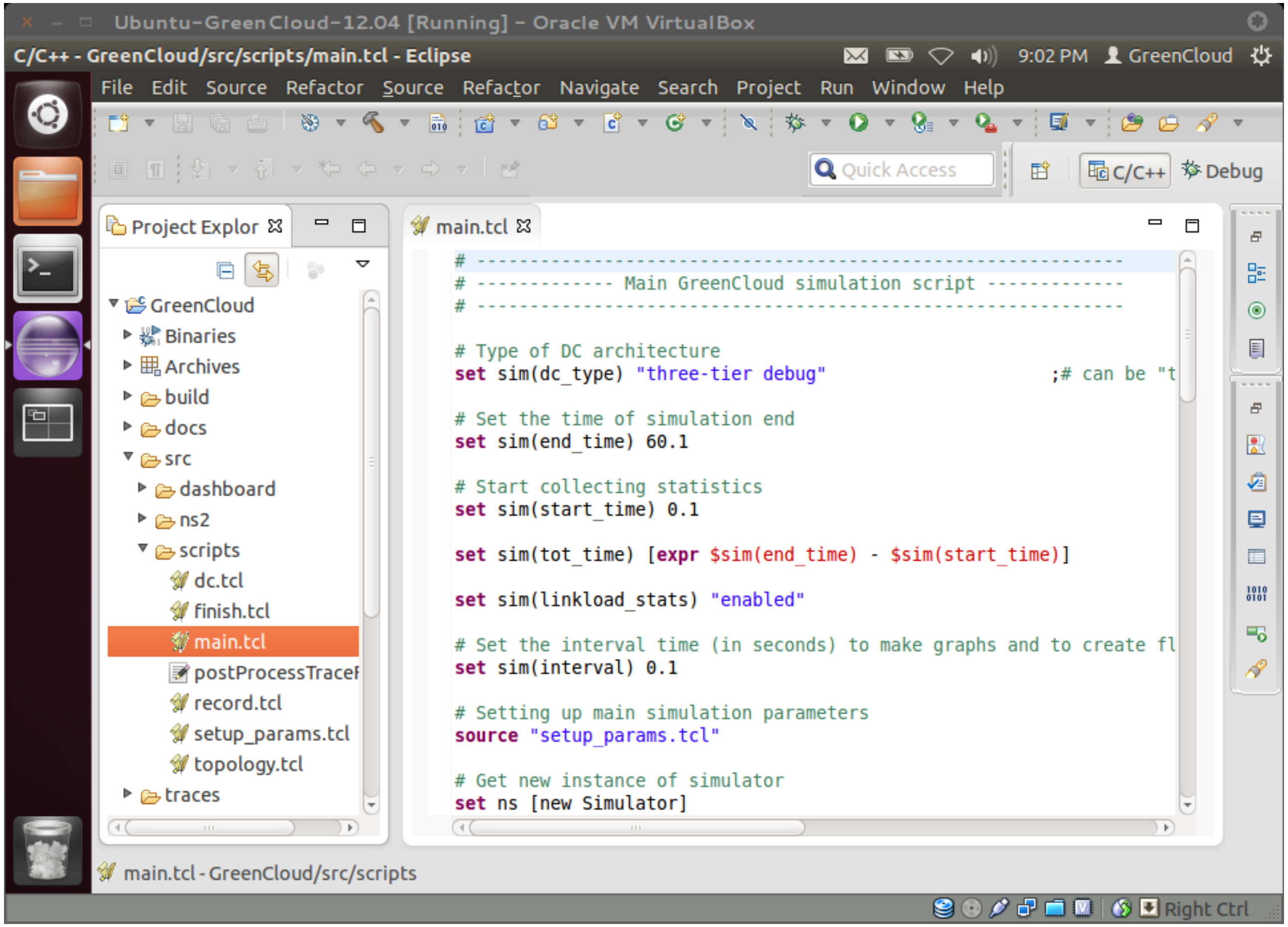
2: Applying optimization solutions



3: Validation and proof of concept

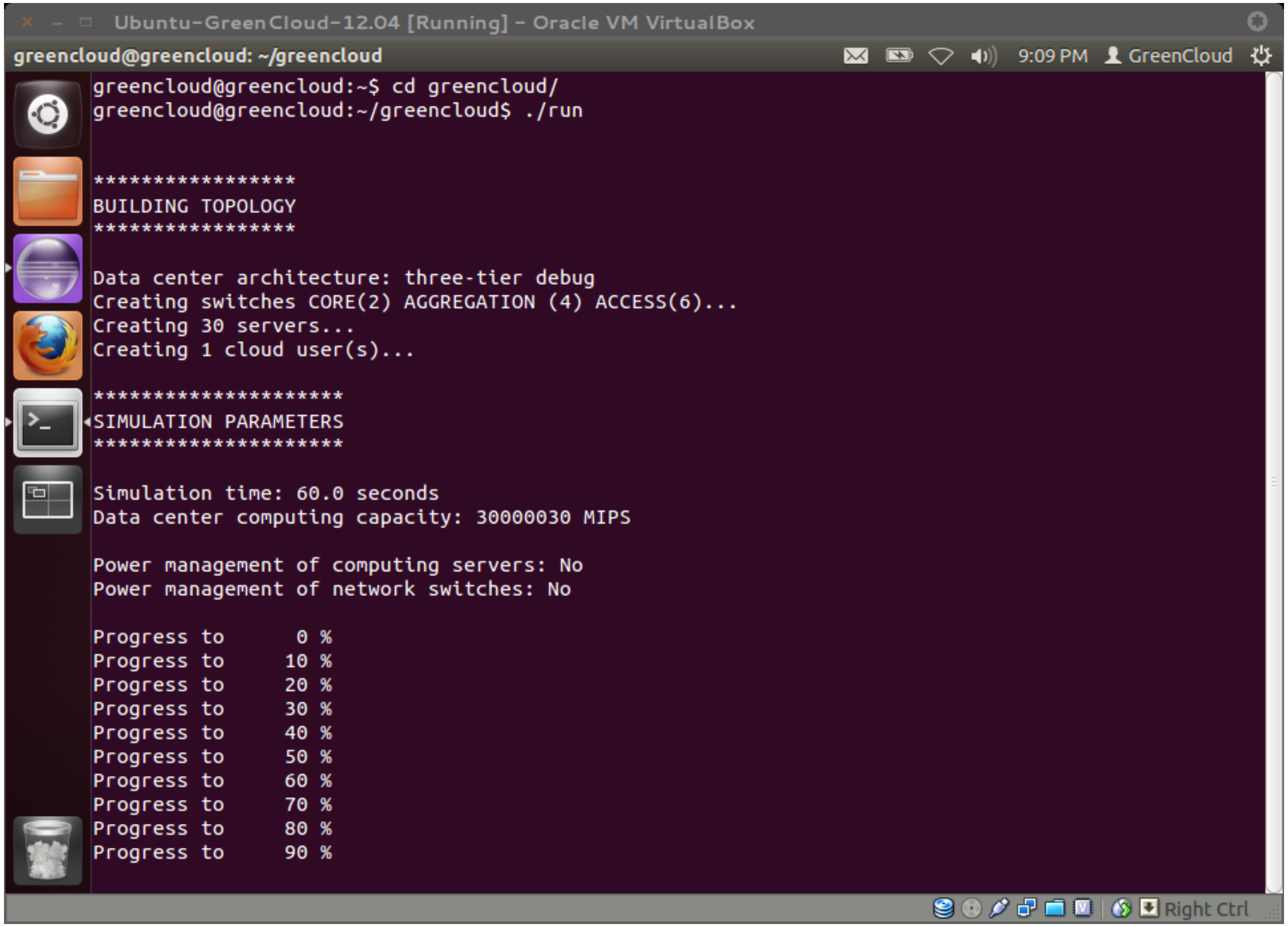
- Can be used to
  - Optimize existing data centers
  - Guide capacity extension decisions
  - Help to design future data center facilities

GreenCloud: Screenshots



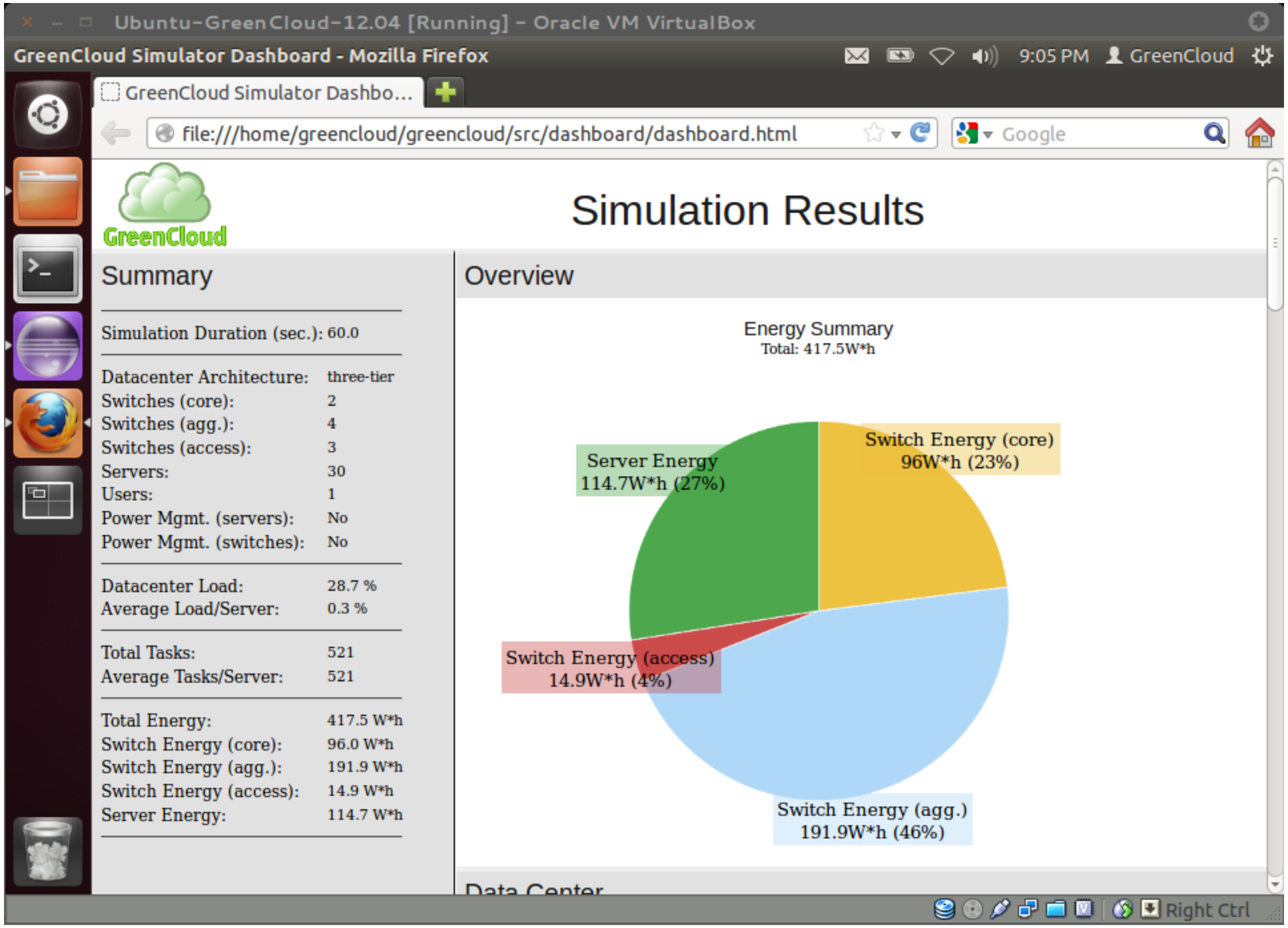
UNIVERSITÉ DU LUXEMBOURG

GreenCloud: Screenshots



UNIVERSITÉ DU LUXEMBOURG

GreenCloud: Screenshots



Modeling Cloud Computing Applications

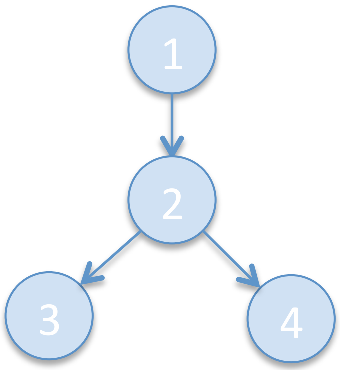
CA-DAG

Pascal Bouvry (pascal.bouvry@uni.lu)

UNIVERSITÉ DU LUXEMBOURG

Modeling of Cloud Applications

- Directed Acyclic Graphs (DAGs)
  - Vertices represent computing tasks of a job
  - Edges represent task dependencies and order of execution



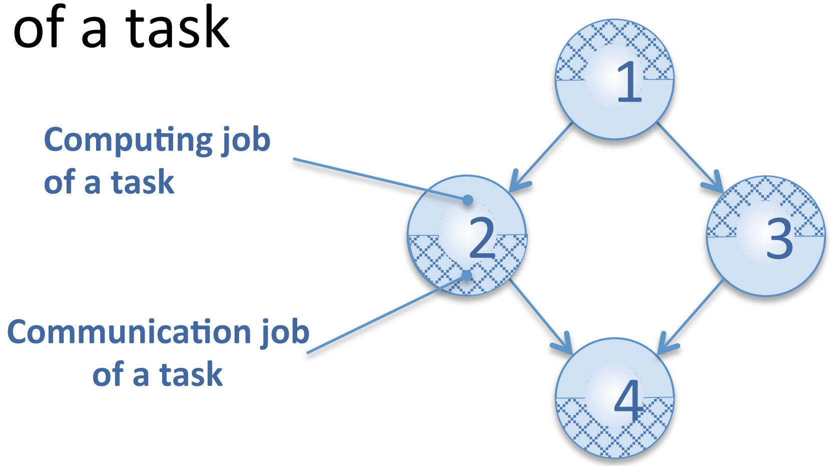
How to model communication processes?

Modeling of Cloud Applications

- Communication-unaware model
- Edges-based model

Modeling of Cloud Applications

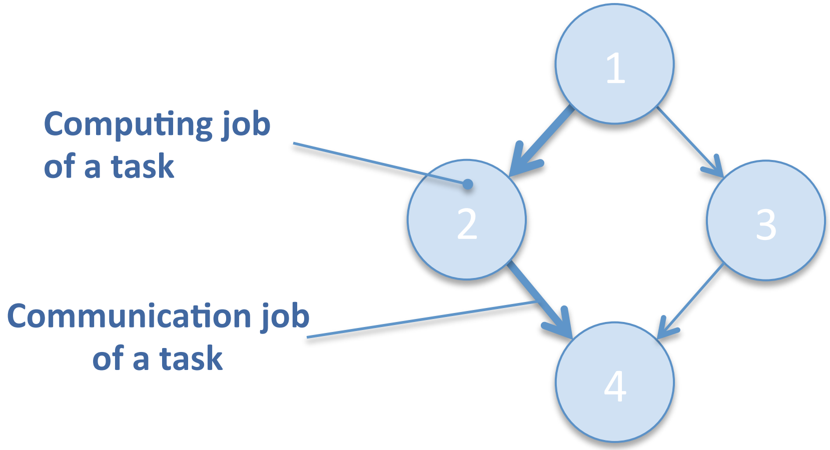
- Communication-unaware model
  - Each vertex represents both computing and communication processes of a task



- Main drawback
  - Having a single vertex for both computing and communications makes it impossible to make separate scheduling decisions

Modeling of Cloud Applications

- Edge-based model
  - DAG edges represent communication processes of a task

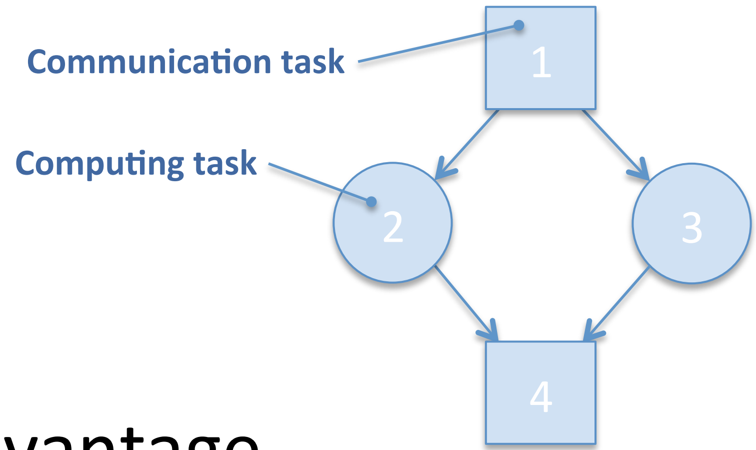


- Main drawback
  - Two different computing tasks cannot have the same data transfer to receive input as a singe edge cannot lead to two different vertices

Proposed Communication-Aware DAG model

Modeling of Cloud Applications

- Proposed CA-DAG: Communication-Aware DAG model
  - Two types of vertices: one for computing and one for communications
  - Edges show define dependences between tasks and order of execution



- Main advantage
  - Allows separate resource allocation decisions, assigning processors to handle computing jobs and network resources for information transmissions

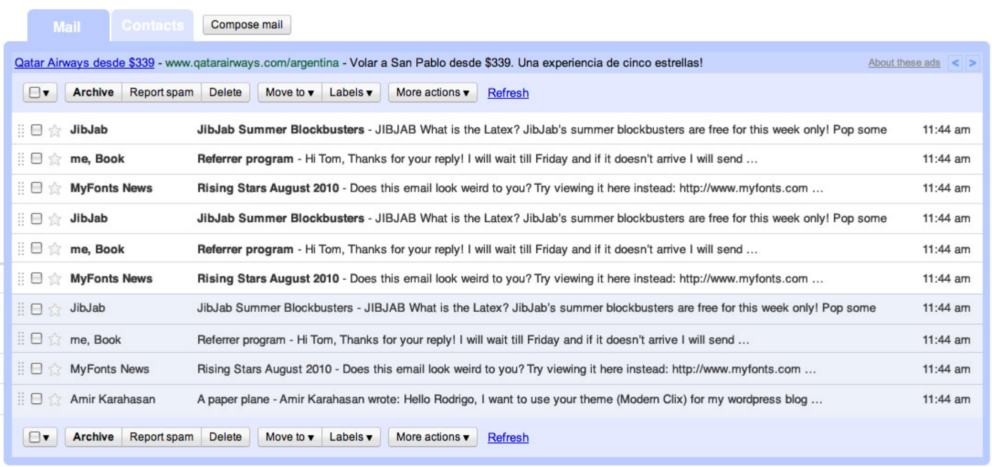
Modeling of Cloud Applications

- Proposed CA-DAG: Communication-Aware DAG model
  - Represented by a directed acyclic graph
  - Set of vertices is composed of computing tasks and communication tasks
  - A **computing task** is described by a pair with the number of instructions (amount of work) that has to be executed within a specific deadline
  - A **communication task** is described by parameters and defined as the amount of information in bits that has to be successfully transmitted within a predefined deadline
  - The set of edges consists of directed edges representing dependence between node and node

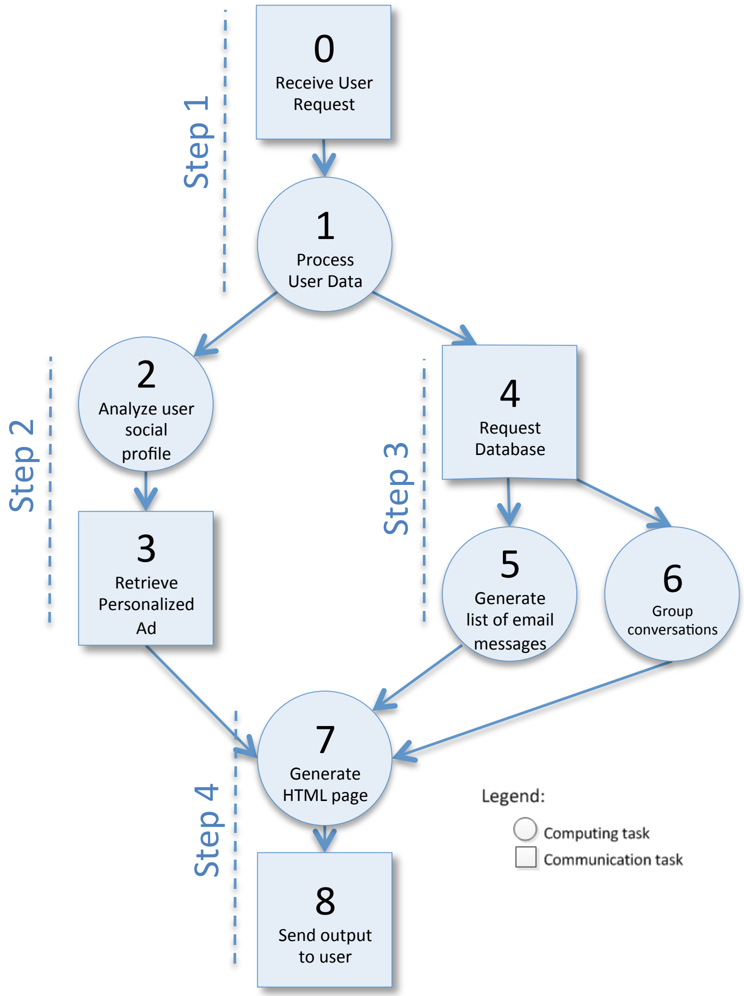
UNIVERSITÉ DU LUXEMBOURG

Modeling of Cloud Applications

- Example of webmail cloud application
  - Step 1: Receive user request and process it
  - Step 2: Generate personalized advertisement
  - Step 3: Request list of email messages from database
  - Step 4: Generate HTML pages and send it to the user

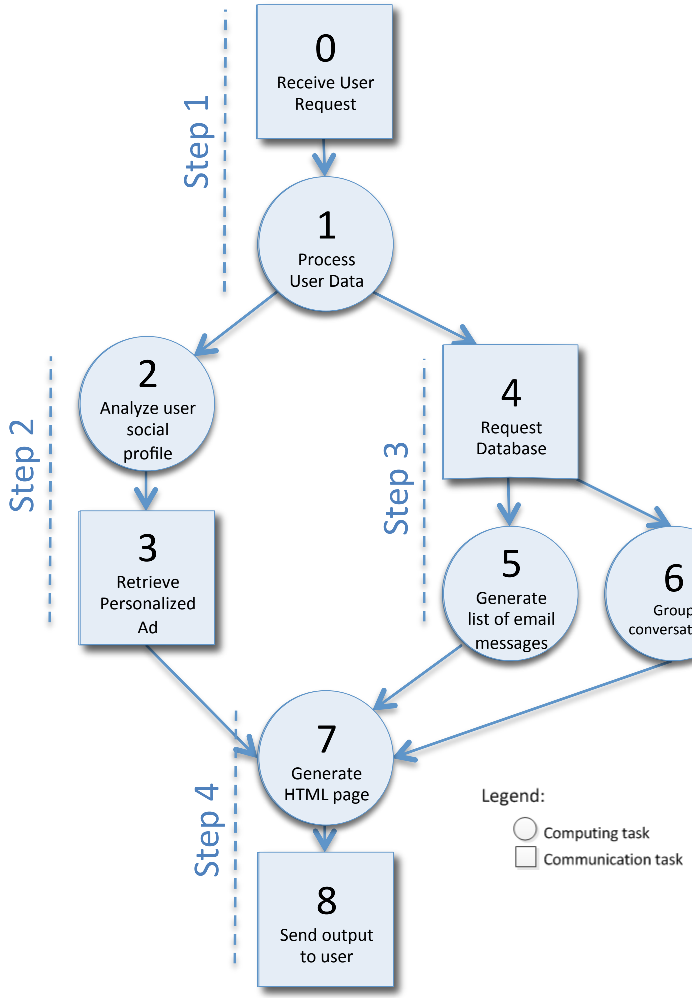


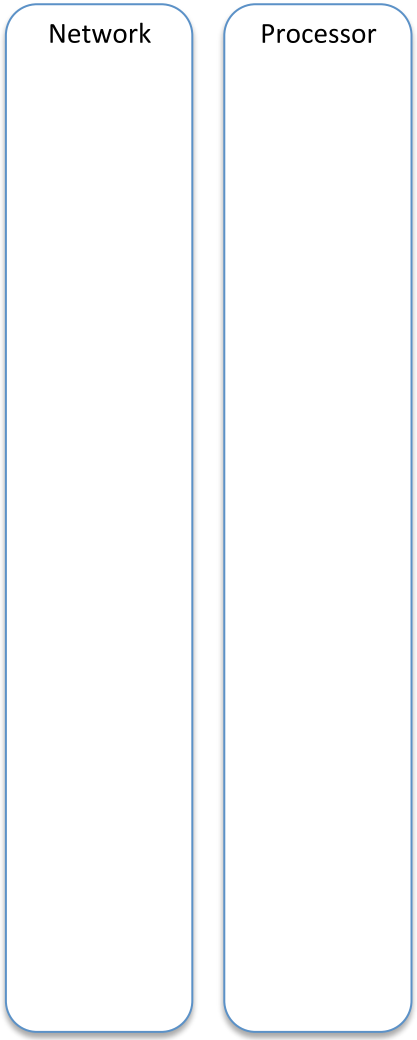
CA-DAG: Communication-Aware DAG



- Step 1: Receive user request and process it
- Step 2: Generate personalized advertisement
- Step 3: Request list of email messages from database
- Step 4: Generate HTML pages and send it to the user

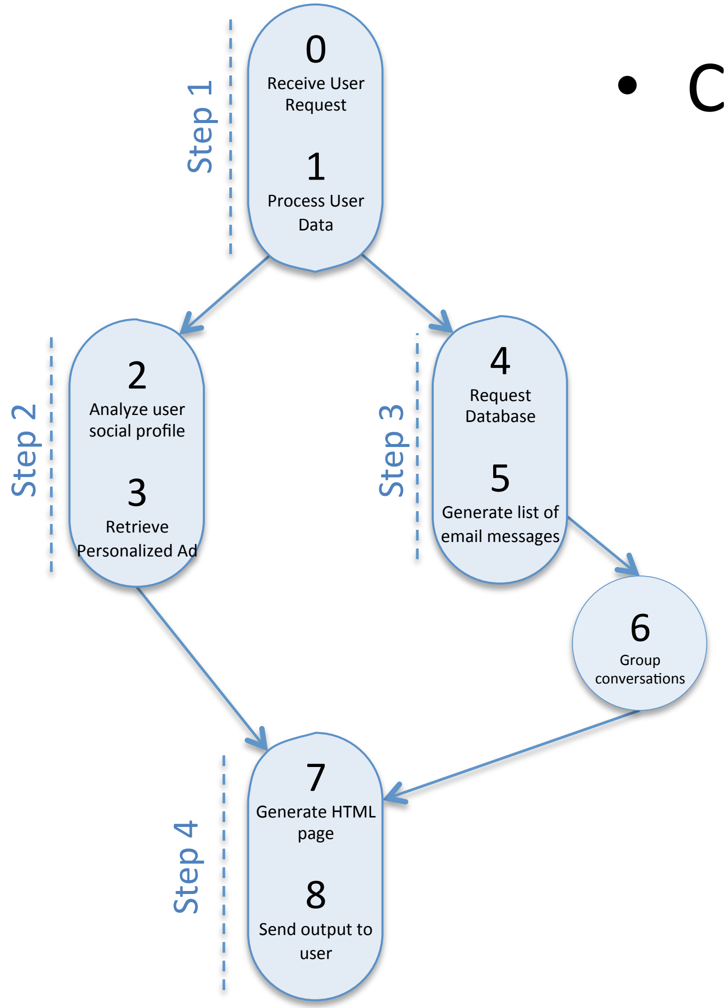
CA-DAG: Communication-Aware DAG





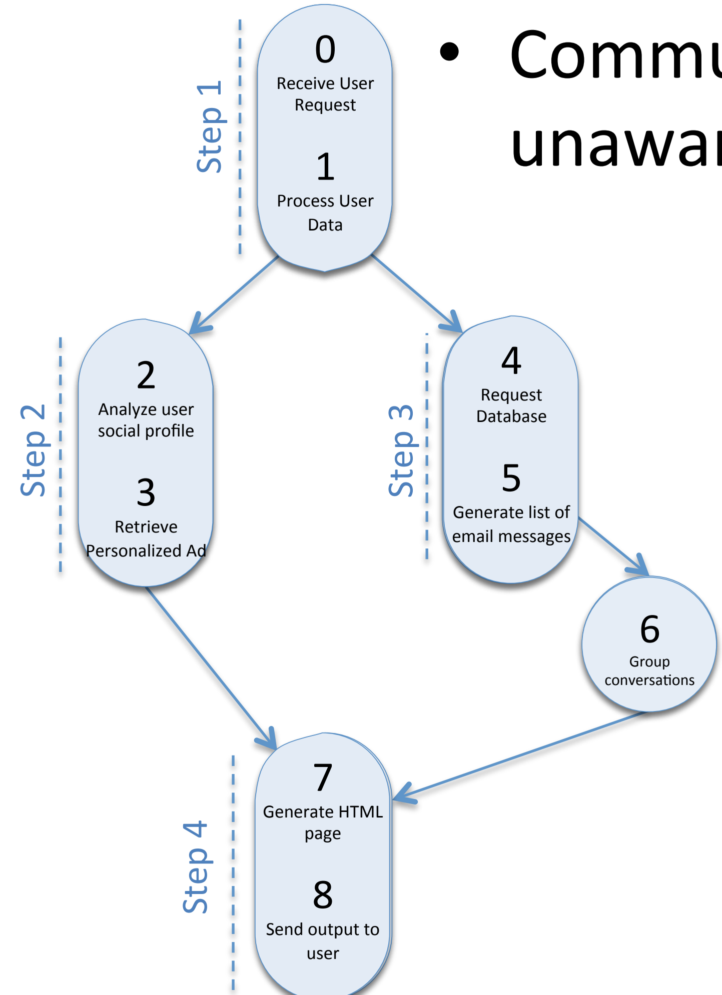
Makespan: 7

CA-DAG: Communication-Aware DAG

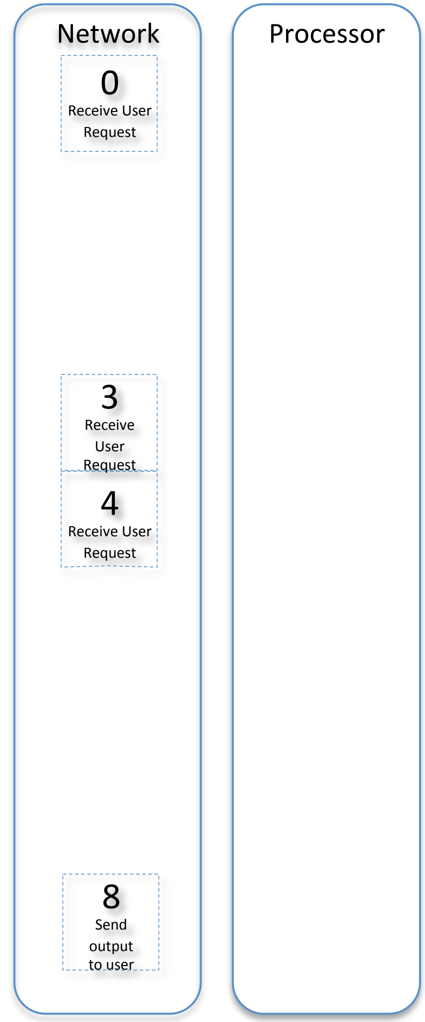


- Communication unaware model
  - Step 1: Receive user request and process it
  - Step 2: Generate personalized advertisement
  - Step 3: Request list of email messages from database
  - Step 4: Generate HTML pages and send it to the user

CA-DAG: Communication-Aware DAG

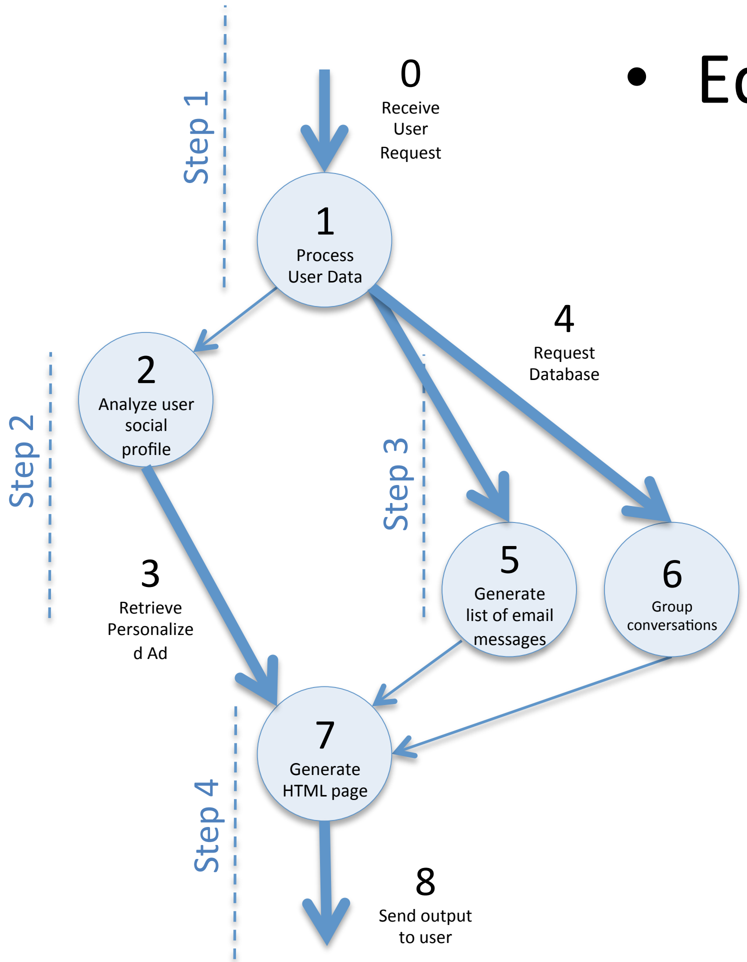


- Communication unaware model



Makespan: 9

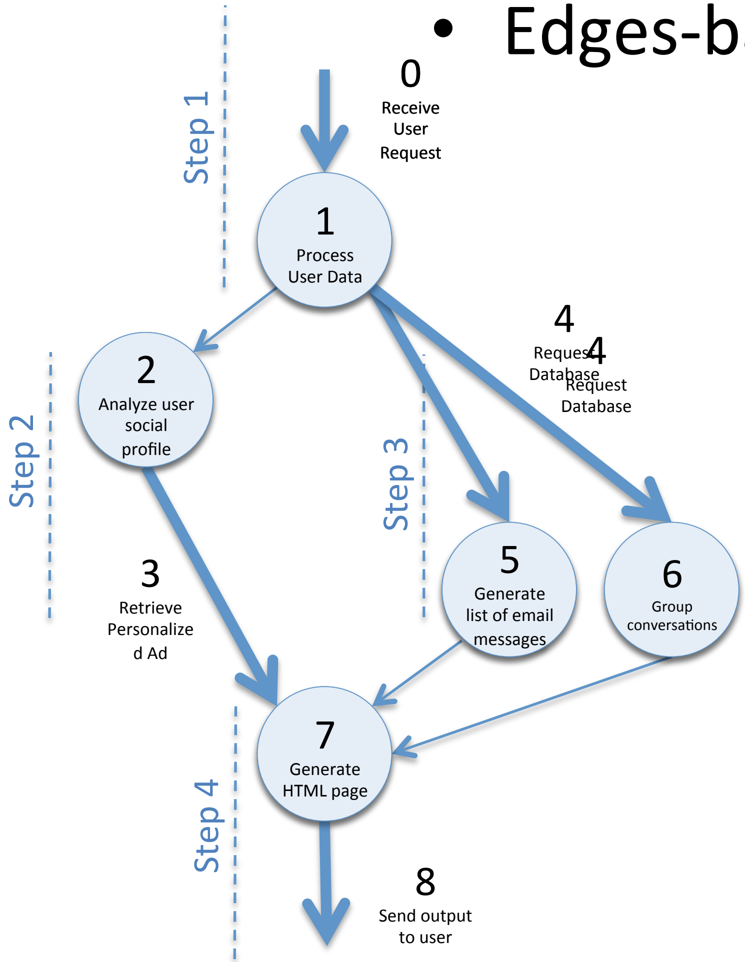
CA-DAG: Communication-Aware DAG

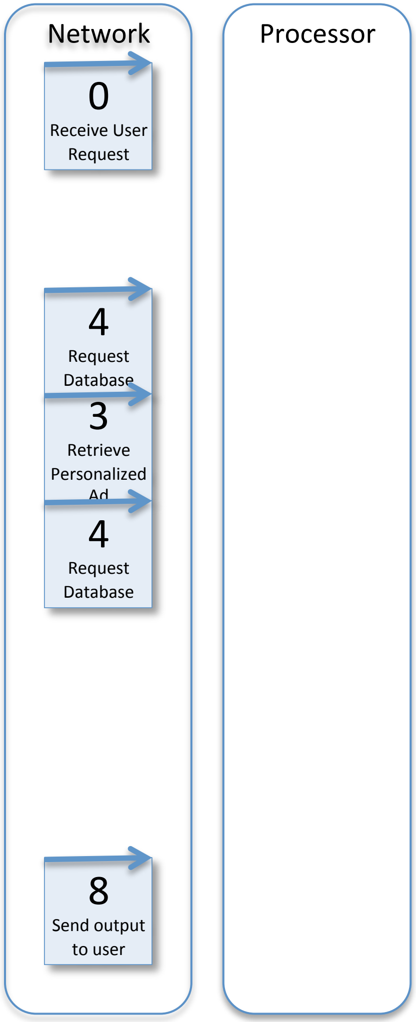


- Edges-based model
  - Step 1: Receive user request and process it
  - Step 2: Generate personalized advertisement
  - Step 3: Request list of email messages from database
  - Step 4: Generate HTML pages and send it to the user

CA-DAG: Communication-Aware DAG

- Edges-based model



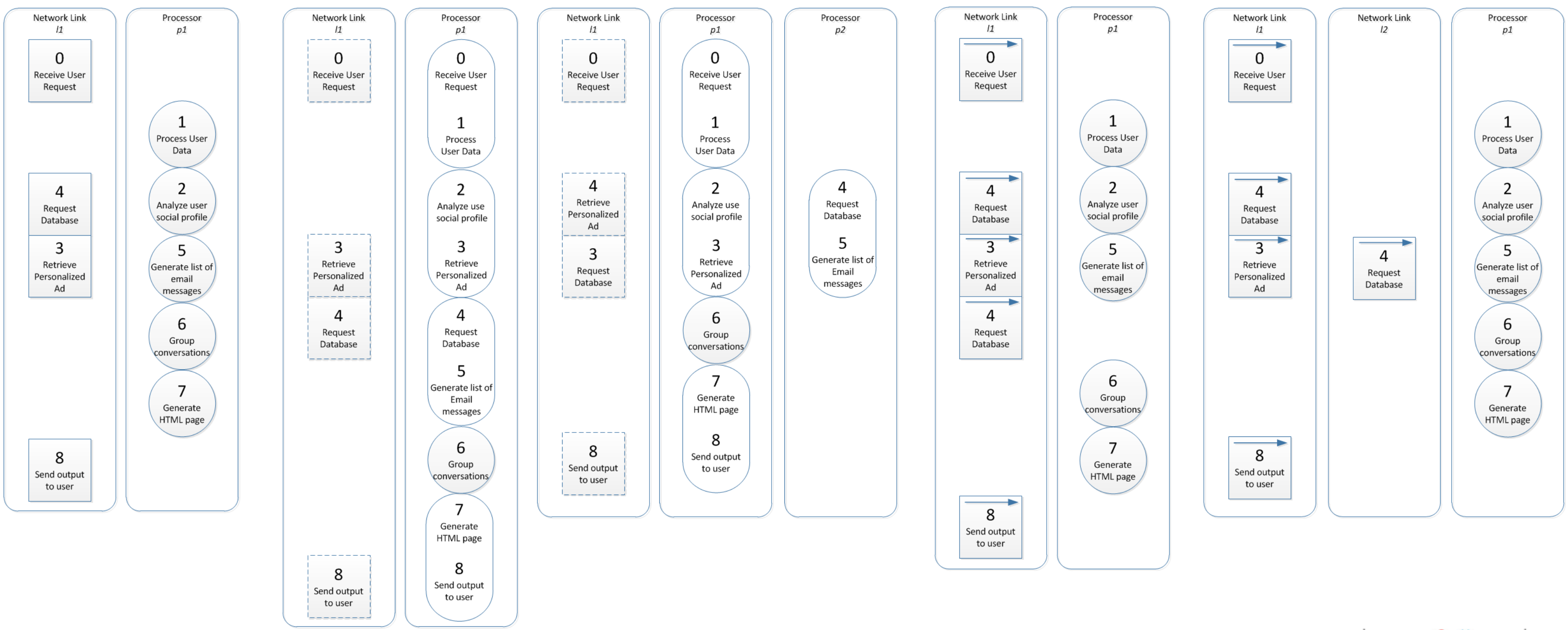


Makespan: 8

CA-DAG: Communication-Aware DAG

- Comparison of schedules

CA-DAG model **Communication-unaware model** **Edges-based model**



CA-DAG: Communication-Aware DAG

- Comparison of models' makespan

# of Processors	# of Network links	Communication-unaware model	Edges-based model	**Proposed CA-DAG model**
1	1	9	8	**7**
1	2	9	7	**7**
2	1	7	8	**7**

Achieves minimum makespan with the least resources

UNIVERSITÉ DU LUXEMBOURG

Properties of Communication Tasks/Vertices

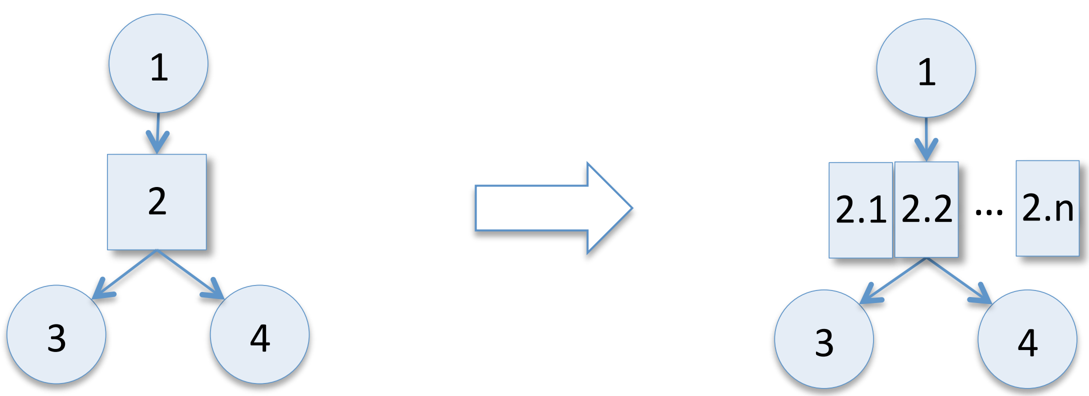
Properties of Communication Tasks/ Vertices

- Task parallelization
- Multipath routing
- Task completion time
- Available bandwidth

Task Parallelization

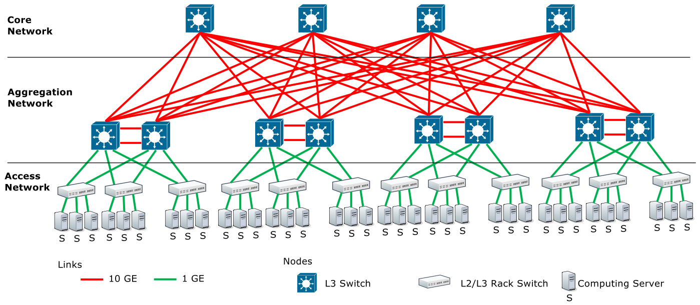
- Each communication task/vertex can be divided into different independent communication tasks that can be executed in parallel

- The smallest size of communication task is one bit as all bits in the message are independent



Multipath Routing

- Most of existing solutions rely on static network topology and fixed pre-allocation which implies circuit switching and pre-defined routing



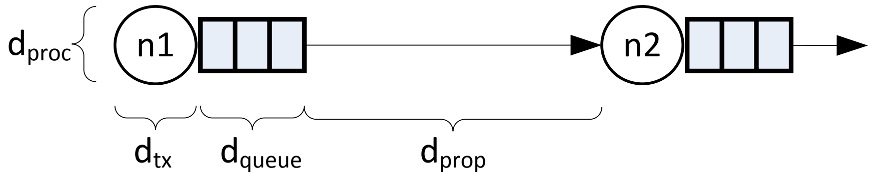
- In reality, datacenter networks are packet switched with routing decisions taken at every hop

- The availability of multiple paths is essential to benefit from parallelization property of communication tasks

Task Completion Time

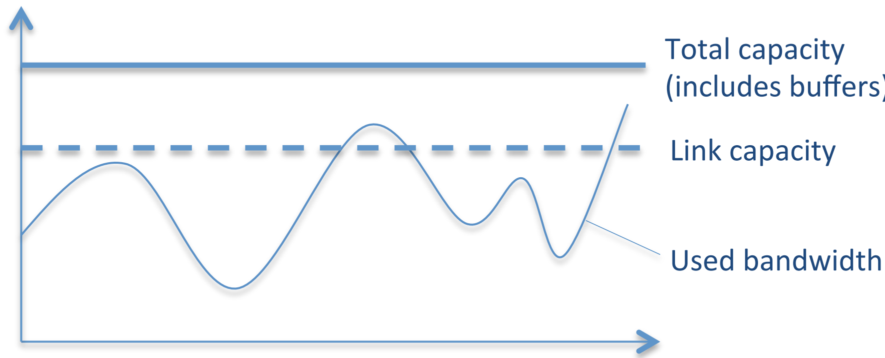
- Execution of communication task involves
  - Packet transmissions on multiple links
  - Sequential processing, variable bitrates
- Communication delay components
  - Processing delay
  - Queuing delay
  - Transmission delay
  - Propagation delay

$$d_{comm} = \sum_{i=1}^{N} (d^i_{proc} + d^i_{queue} + d^i_{tx} + d^i_{prop}).$$



Available Bandwidth

- Residual Bandwidth
  - Bandwidth left unoccupied



- Available bandwidth
  - Bandwidth that a new flow can obtain (residual bandwidth + portion of the used bandwidth)

- Utilization performance of communication protocols
  - TCP throughput

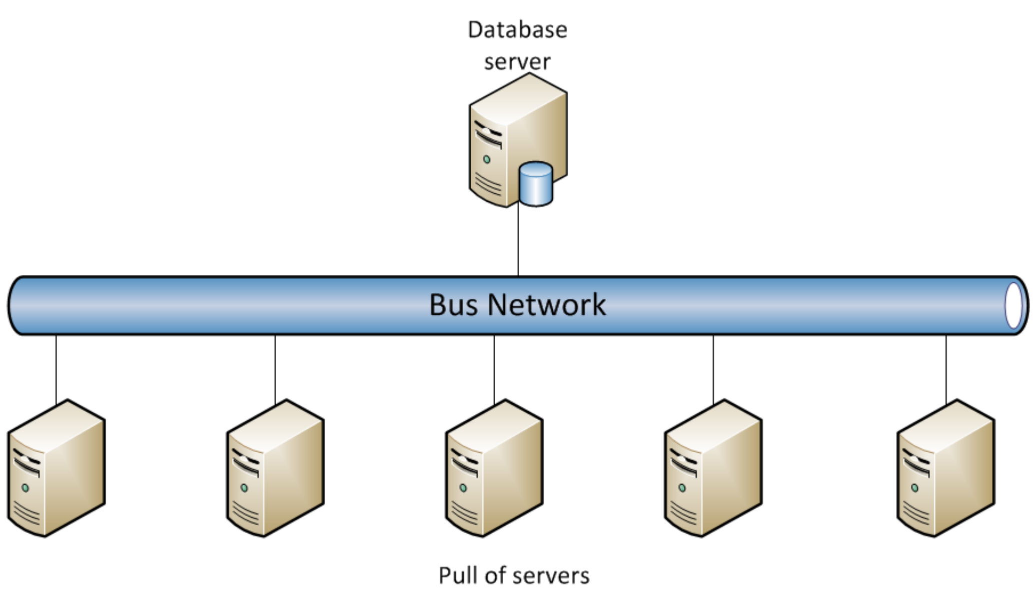
$$B(p) = \frac{MSS}{RTT \cdot \sqrt{p}},$$

Efficiency of CA-DAG Model

UNIVERSITÉ DU LUXEMBOURG

Efficiency of CA-DAG model

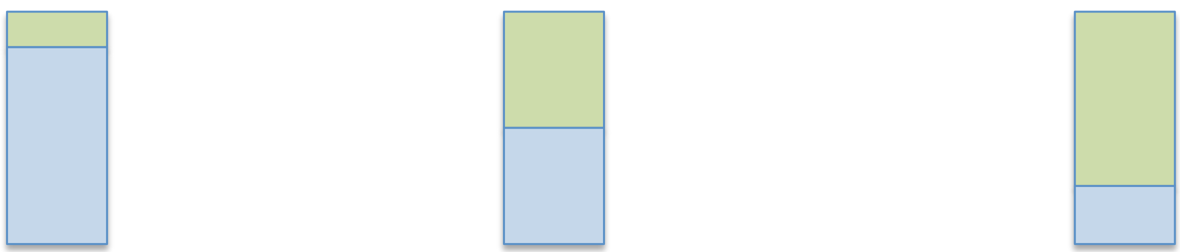
- System Architecture
  - Only one node can communicate at a time



Efficiency of CA-DAG model

- Workloads
  - Winkler graph generator
  - DAGs with occasional and frequent communications

- Communication-to-Computation Ratio (CCR)



0.1
Computationally intensive DAGs

1
Balanced DAGs

2
Communication intensive DAGs

Efficiency of CA-DAG model

- Scheduling Algorithm
  - Offline (deterministic) scheduling
  - Zero release time of DAGs
  - Clairvoyant execution and communication time
  - Adapted list scheduling is employed
  - A processor allowing minimum execution time is selected

Efficiency of CA-DAG model

- Scheduling Criteria
  - Schedule efficiency
  - Approximation factor

Efficiency of CA-DAG model

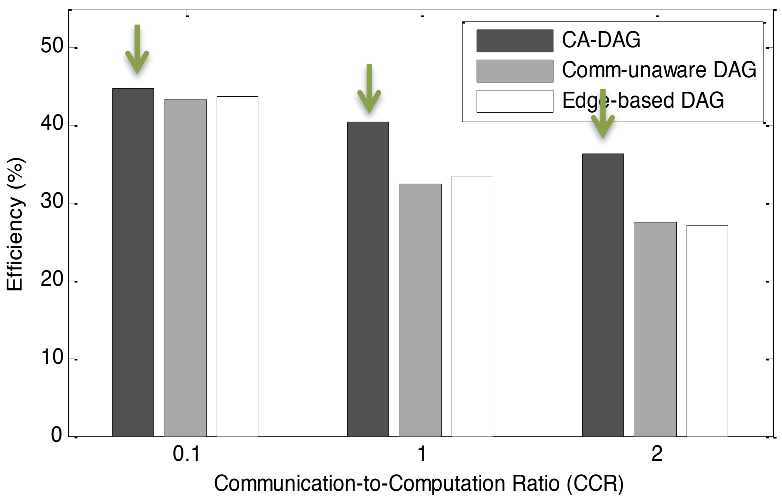
Schedule efficiency:

Ratio of sequential execution time to the makespan by the number of computing resources

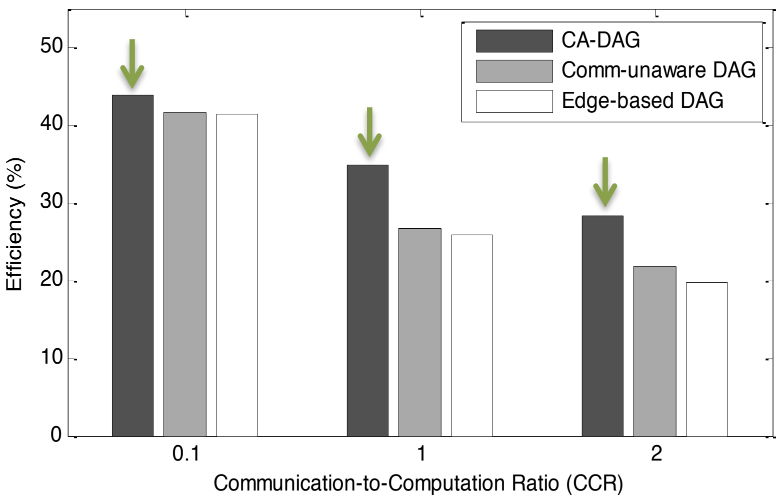
$$\mathrm{eff}(S) = \frac{\sum_{i=1..,n}(p_i)}{C_{max} \times m}$$

Efficiency of CA-DAG model

- Schedule efficiency
  - Apps. with occasional communications
  - Apps. with frequent communications



The higher the better



Efficiency of CA-DAG model

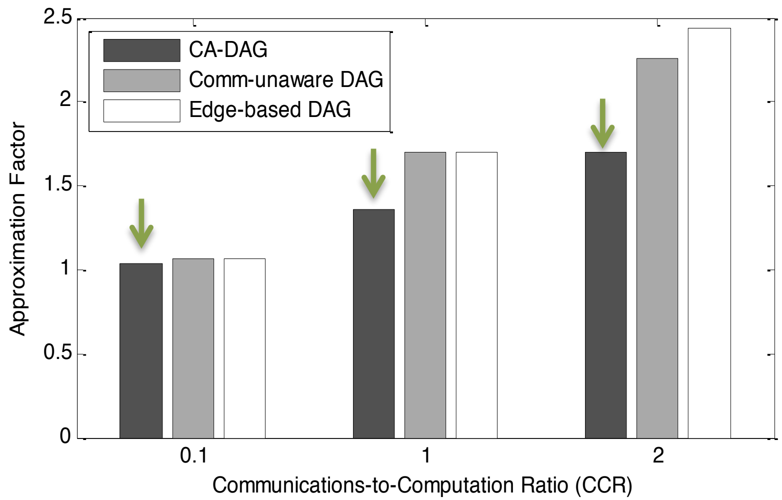
$$\rho = \frac{C_{\max}}{C^*_{\max}}$$

$$C^*_{max} \geq \tilde{C}^*_{max} = max\left\{\max(blevel(t_i))\,,\ \frac{\sum_{i=1..,n}(p_i)}{m}\right\}$$

Efficiency of CA-DAG model

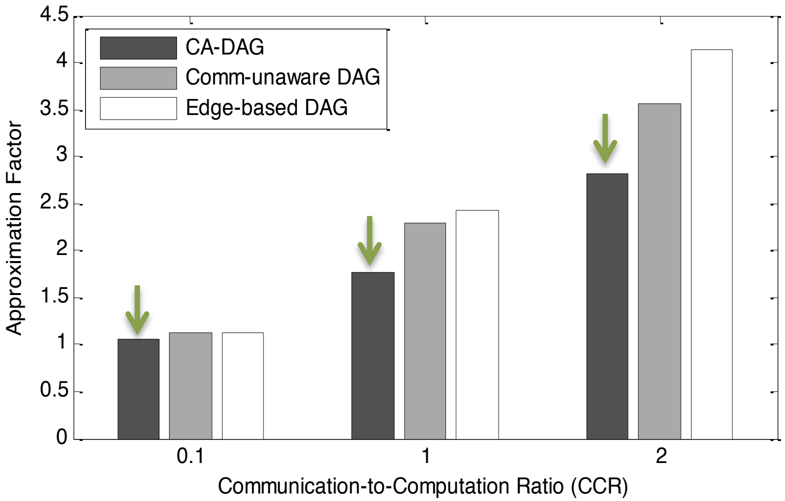
- Approximation factor

  - Apps. with occasional communications



The lower the better

  - Apps. with frequent communications



UNIVERSITÉ DU LUXEMBOURG

CA-DAG model

- Cloud applications use communication resources excessively
- New communication-aware model of cloud applications, named CA-DAG, is proposed
- CA-DAG includes separate vertices to represent communication processes to allow making separate resource allocation decisions (computing jobs to processors, communication jobs to the network)
- CA-DAG model enables the design of novel solutions with mixed scheduling policies optimized for cloud computing

Conclusion

- We have
  - Benchmarked classical hypervisors
  - Highlighted the communication issue
  - Proposed a new cloud simulator called Greencloud
  - Proposed an enhanced DAG model called CA-DAG

- Cloud computing is there but comes at a cost.

UNIVERSITÉ DU LUXEMBOURG

Perspectives and other aspects

- New generations of VMs and HPC PaaS
- Network coding
- Hybrid cloud (public/private) solutions and cloud brokering
- New generation HW, mixing ARMs and GPUs
- Legal and security aspects

Thank you!

Contact information:

Pascal Bouvry

University of Luxembourg

Pascal.Bouvry@uni.lu

UNIVERSITÉ DU LUXEMBOURG