Virtualization experience with Xen

Havard Bjerke

02.05.2006

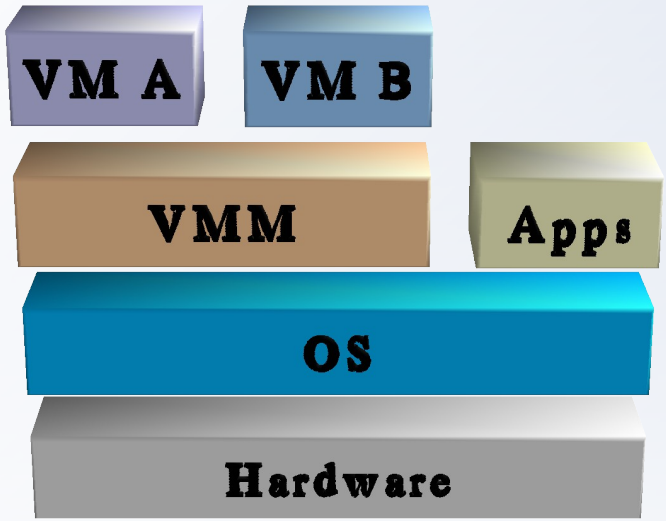
CERN
openlab for DataGrid applications
Developing Solutions for the Data-Intensive Science of the Large Hadron Collider

Overview

- VM technology
- Openlab I
  - History
  - Xen/ia64
  - Deployment in the LCG testbed
- Openlab II
  - New hardware
  - ETICS, Smartfrog
  - Virtualization in batch subsystem
- Vision

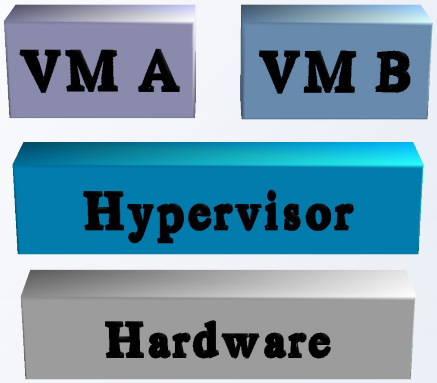
Virtualization Technologies - Hosted

- Microsoft Virtualization Server
  - Used at CERN for consolidation
    - Runs MS Windows, Linux (SLC3, SLC4)
    - Non-negligible CPU overhead: every ~ 3rd cycle wasted
  - Hosted
  - Free
  - API to control VMM and VMs
  - 32-bit, single CPU only
- VMWare
  - Hosted: GSX
    - Non-negligible CPU overhead
  - Non-hosted: ESX
    - Limited hardware support



Virtualization Technologies - Non-hosted

- Xen
  - Paravirtualization, non-hosted
  - Close to native performance
  - Supports only paravirtualized OSs unless hardware-virtualized platform
  - 64-bit support
  - SMP support
  - Open source, GPL



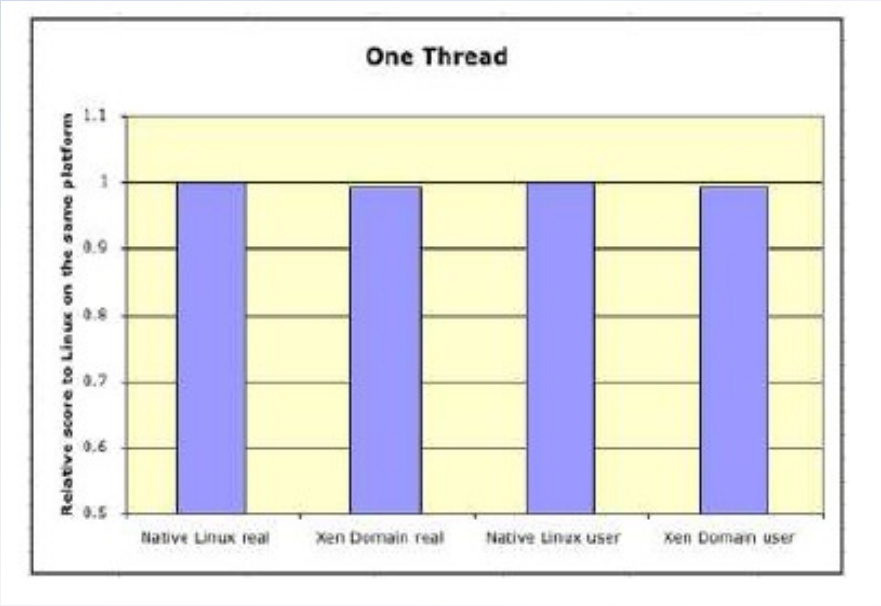
Openlab I

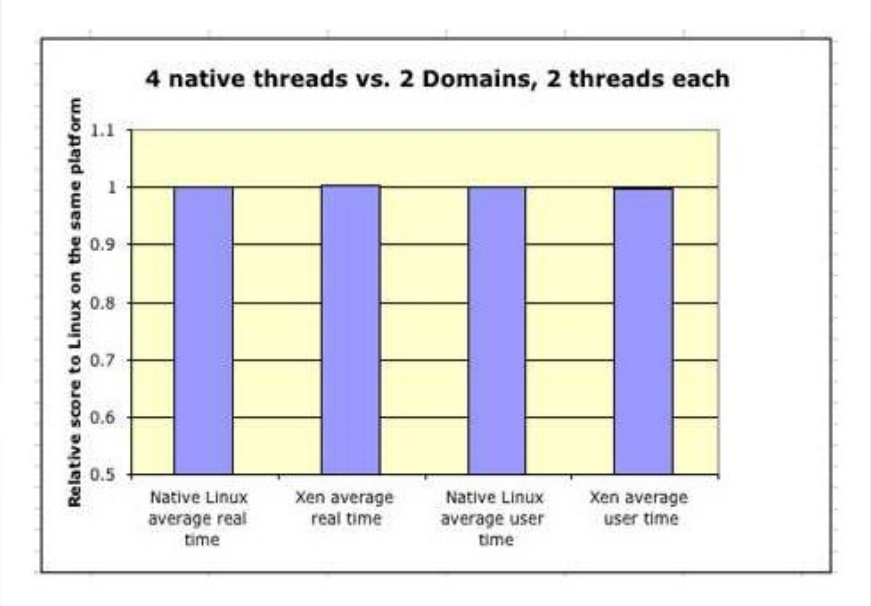
Havard Bjerke

02.05.2006

CERN
openlab for DataGrid applications
Developing Solutions for the Data-Intensive Science of the Large Hadron Collider

CPU performance benchmarks (Rune)





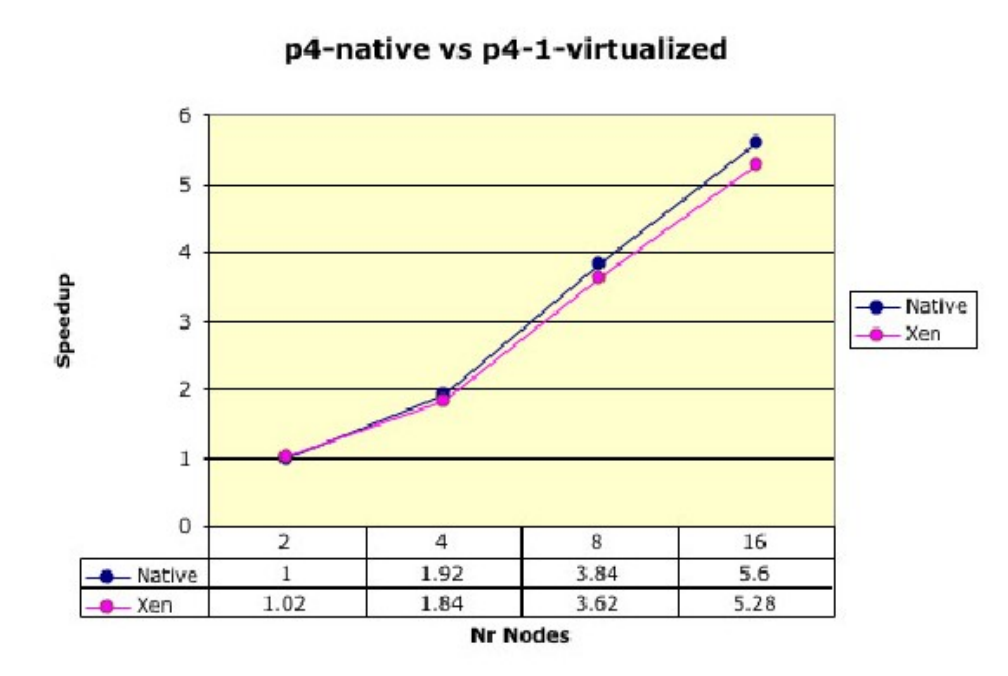
CPU: Dual Xeon 2.4 Ghz
Benchmark: test40forSPEC
OS: SLC3

Cluster benchmarks (with Rune)

- Lower bandwidth and higher latency in guest domains.
- Aggregate bandwidth of multiple domains scales, but not latency.

Configuration	Bandwidth	Startup
p4-native	116.60 MByte/s	119.96 μs
p4-1-virtualized	100.43 MByte/s	150.89 μs

Configuration	Bandwidth	Startup
2-partitioned (A)	50.81 MByte/s	4228 μs
2-partitioned (B)	51.10 MByte/s	4247 μs

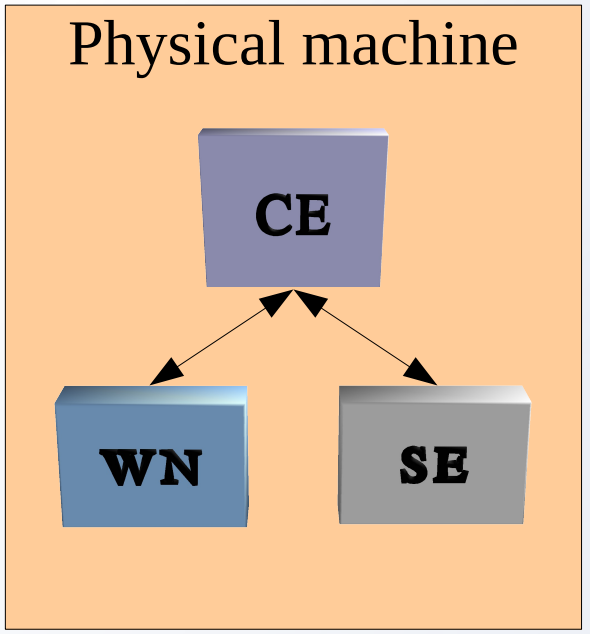


Xen/ia64

- X86 virtualization unfriendly
  - Non-faulting privilege sensitive instructions
- IA64 a little more friendly
  - Three non-faulting privilege sensitive instructions
  - Tagged TLB / Region registers
    - No need to flush the TLB when switching domain
    - Easier to virtualize physical memory
  - No more segmentation, real mode, protected mode
    - Redundant hypercalls
  - EFI – easier to virtualize than BIOS
- Optimized paravirtualization
  - Linux is a moving target -> minimize changes in the guest Linux kernel
  - Instead: trap faulting instructions

LCG Deployment

- Xen 2.0.7
- Proof of concept GRID-in-a-box
- Complete LCG 2.6 installation
  - Computing Element (CE)
  - Storage Element (SE)
  - Worker Node (WN)



LCG Deployment

- Successful deployment in LCG testbed
  - Currently used in test grid
  - Tests passed as good as native nodes
- Issues
  - Automatic software updates cause /lib/tls to be restored
  - Support for 2.4 kernels dropped
- Possible applications
  - Server consolidation (GRID-in-a-box)
  - Security (VO-box)
  - Customizable environments (Openlab II)
  - Availability, management flexibility (Live-migration)

Openlab II
Focused effort with Intel

Havard Bjerke

02.05.2006

New Hardware

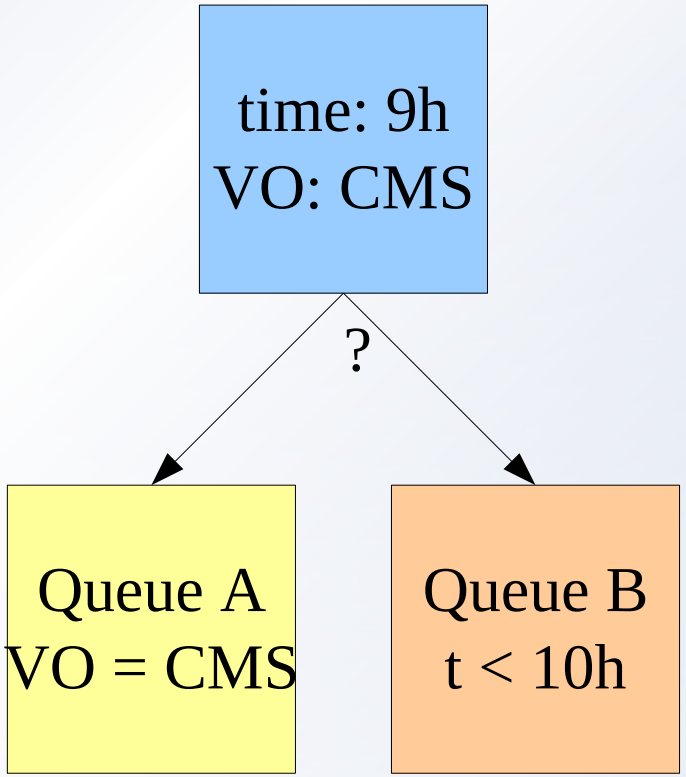
- Montecito
  - IA64
  - VTi
  - Paravirtualization – OK
    - No support for FPSWA yet
- Dempsey
  - x86 EM64T
  - VTx
  - Paravirtualization - OK

Smartfrog, ETICS

- ETICS (Xavier)
  - Testing environment
- Smartfrog
  - Utility computing
  - Provide a single configuration file
    - Memory
    - HD capacity
    - Software configuration
    - ...
  - Deploy a complete site – clean up afterwards

Virtualization in Batch Subsystems

- PBS
  - Resource scheduling independent from queues
- LSF
  - One queue per VO
- BLAHP
  - Common interface to batch subsystems
  - Let LSF do scheduling decisions based on time constraints

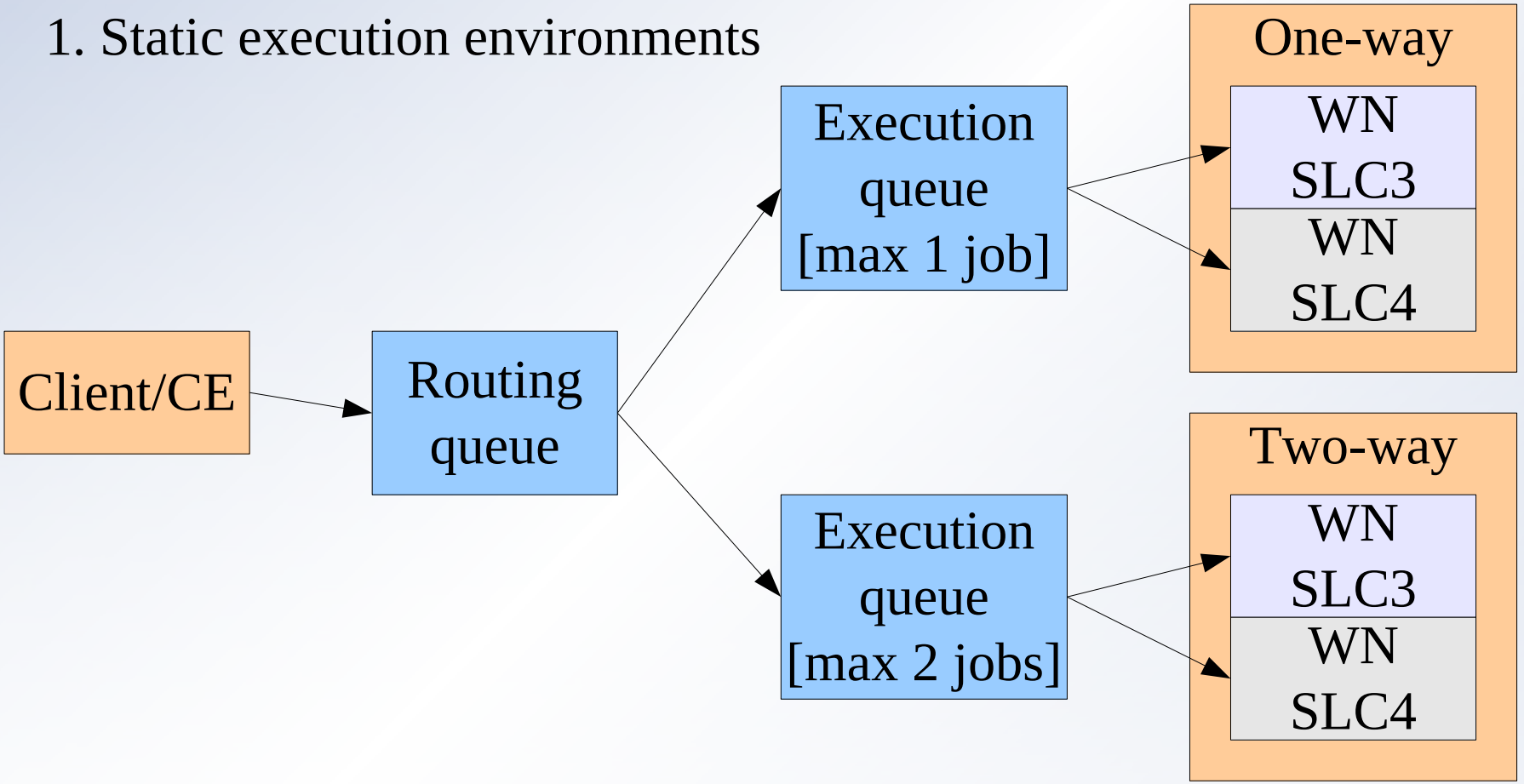


Virtualization in Batch Subsystems

- Requirements
  - Customized execution environments
  - Isolated execution environments
  - Dynamic resource management
- Three goals over three phases
  1. Selection of static execution environments
  2. Dynamic instantiation of execution environments – on-demand
  3. Dynamic configuration of execution environments – VM factory

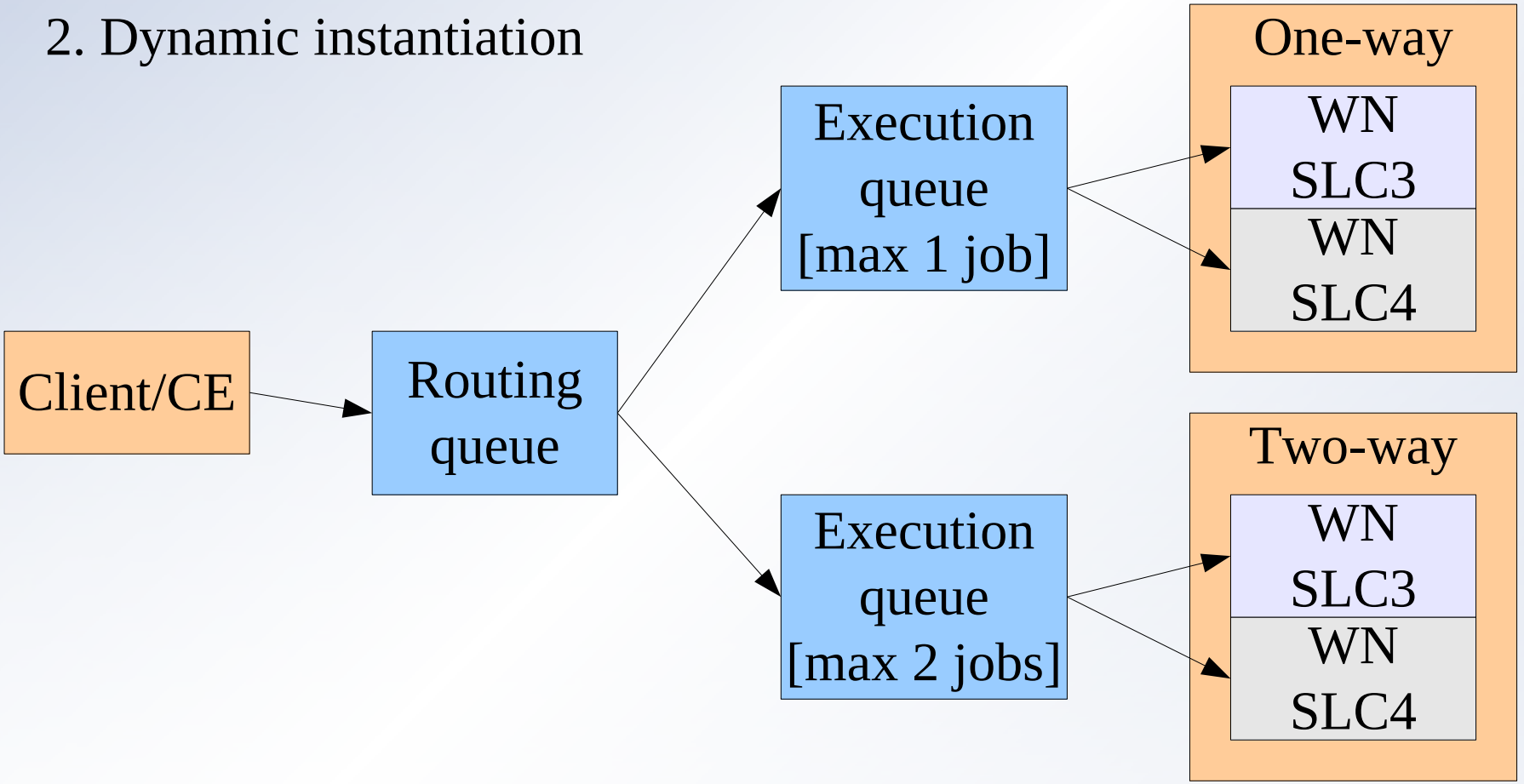
Virtualization in Batch Subsystems

1. Static execution environments



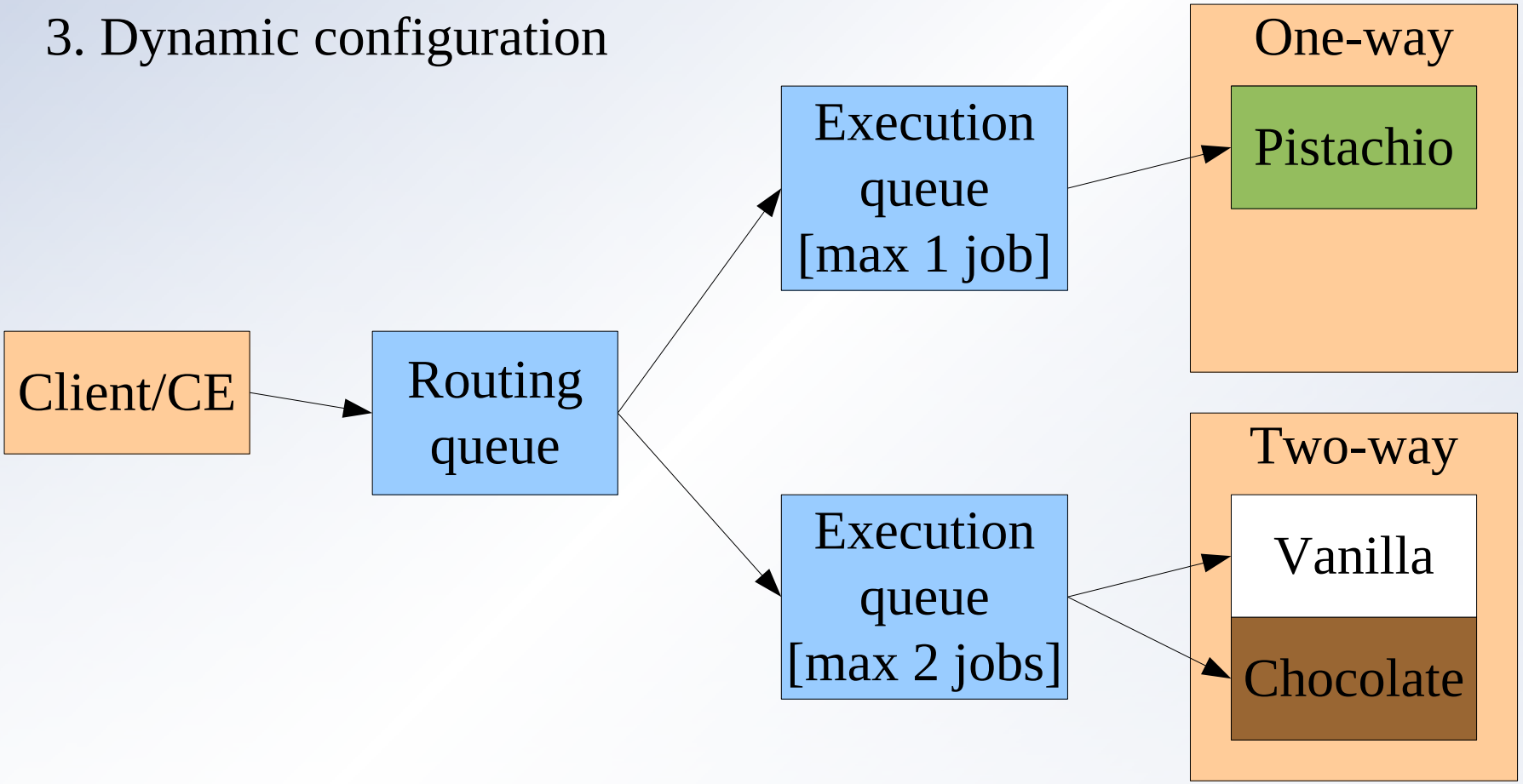
Virtualization in Batch Subsystems

2. Dynamic instantiation



Virtualization in Batch Subsystems

3. Dynamic configuration



Vision

- Close to native performance
  - Without hardware support
    - Xen
    - VMWare ESX
  - With hardware support
    - Many to come
- VM tech agnostic
  - Casatt's XVM
- User-supplied or -specified execution environments
- Domain migration
  - Flexible resource management
  - High availability

Vision

- Execution environment characteristics
  - Isolated
  - Secure
    - XenSE
    - SVM hardware extensions
  - Clean

Questions?

More info:

http://openlab-mu-internal.web.cern.ch/openlab-mu-internal/openlab-II_Projects/Platform_Competence_Centre/Virtualization/Virtualization.asp

Havard Bjerke