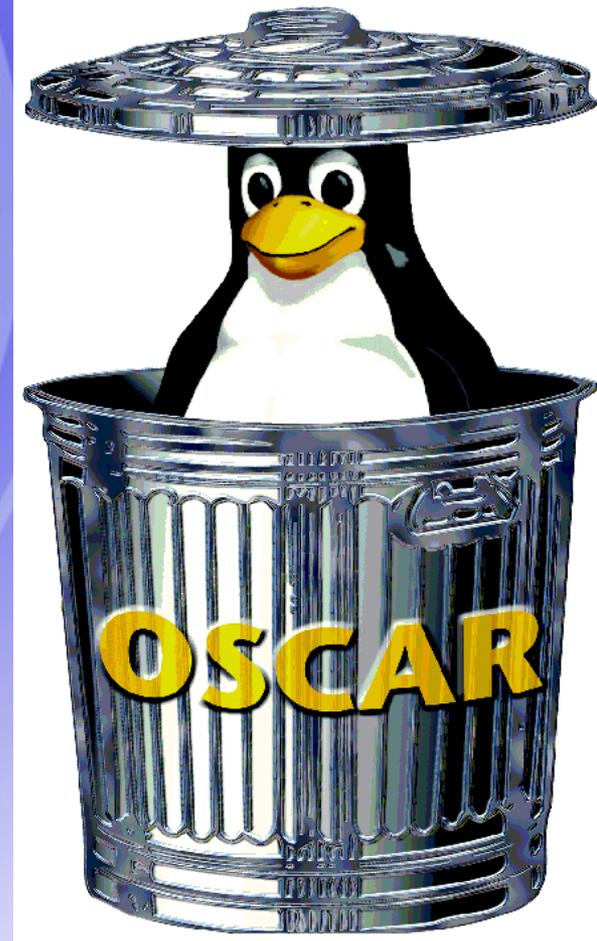


HPCS 2006  
St. John's  
May 2006

# OSCAR-Pro

combining research ideas with  
customer demands

Dr. Erich Focht  
NEC HPC Europe



# Overview

- ◆ What is OSCAR?
  - ◆ Organization, working groups
  - ◆ Functionality and structure
- ◆ What is OSCAR-Pro?
  - ◆ History
  - ◆ Customer examples
  - ◆ Features & roadmap
- ◆ Open Source and academia + industry cooperation

# OSCAR

## Open Source Cluster Application Resources

- ◆ Snapshot of best known methods for building and running clusters for High Performance Computing
  - ◆ leverage wealth of open source components
  - ◆ make building and installing clusters reproducible and easy
- ◆ initiated: January 2000
- ◆ first public release: April 2001
- ◆ project organization
  - ◆ Open Cluster Group (OCG)
  - ◆ consortium of academic/research and industry members
  - ◆ OCG working groups

# OSCAR Member Organizations

## ◆ Academia

OAK RIDGE NATIONAL LABORATORY  
U. S. DEPARTMENT OF ENERGY



pervasivetechlabs  
AT INDIANA UNIVERSITY



Canada's Michael Smith  
G E N O M E  
S C I E N C E S  
C E N T R E



## ◆ Industry



NEC



Bald Guy  
S O F T W A R E

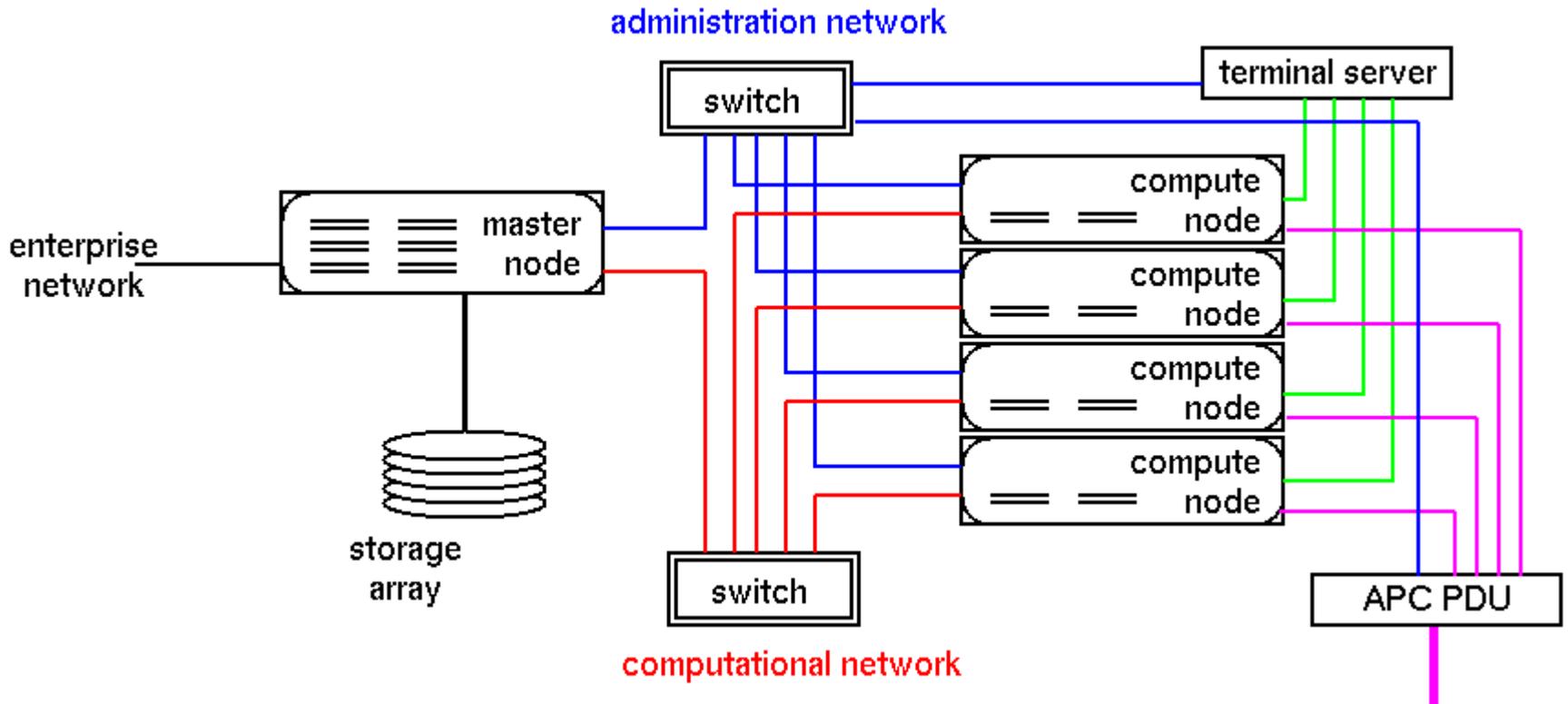
# OSCAR Working Groups

- ◆ OSCAR
- ◆ SSI-OSCAR
  - ◆ single system image: Kerrighed
- ◆ HA-OSCAR
  - ◆ high availability master nodes
- ◆ SSS-OSCAR
  - ◆ scalable systems software

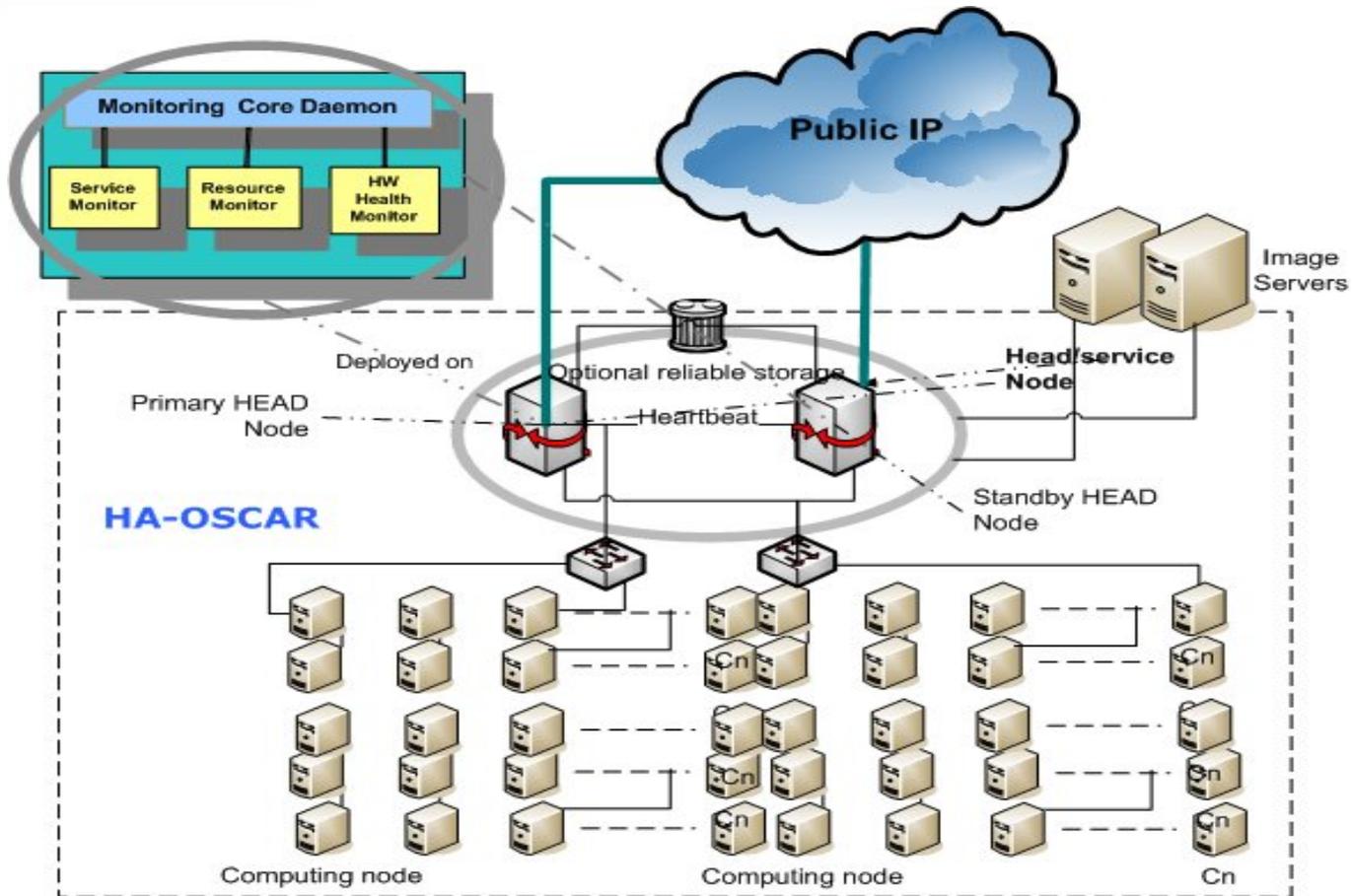
# What does OSCAR do?

- ◆ Wizard based cluster software installation
  - ◆ Operating system (image based deployment)
  - ◆ Cluster environment
- ◆ Automatically configures cluster components
- ◆ Increases consistency among cluster builds
- ◆ Reduces time to build / install a cluster
- ◆ Reduces need for expertise

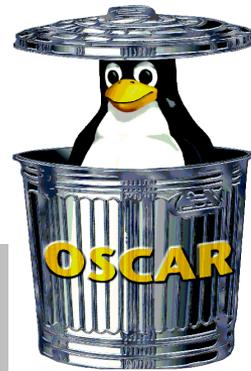
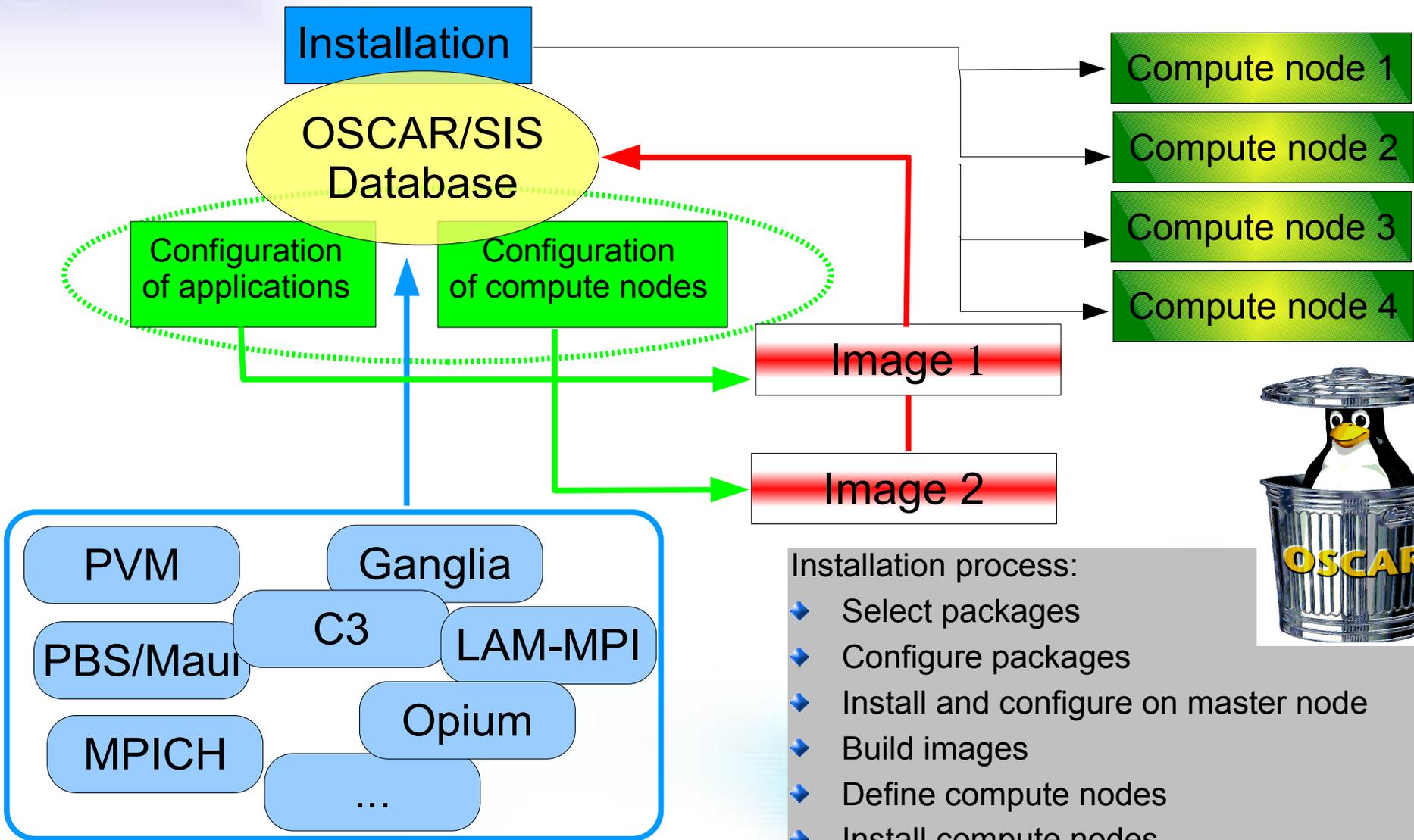
# Standard cluster



# HA-OSCAR Cluster



# OSCAR Structure



## Installation process:

- ◆ Select packages
- ◆ Configure packages
- ◆ Install and configure on master node
- ◆ Build images
- ◆ Define compute nodes
- ◆ Install compute nodes
- ◆ Configure packages on whole cluster

# What is OSCAR-Pro ?

- ◆ OSCAR
  - ◆ NEC specific add-ons
  - ◆ Services
  - ◆ Customer oriented R&D
- 
- ◆ Linux cluster solution deployed by NEC HPC Europe to its customers.

# NEC HPC Europe: Offices

Düsseldorf / Germany

Stuttgart / Germany

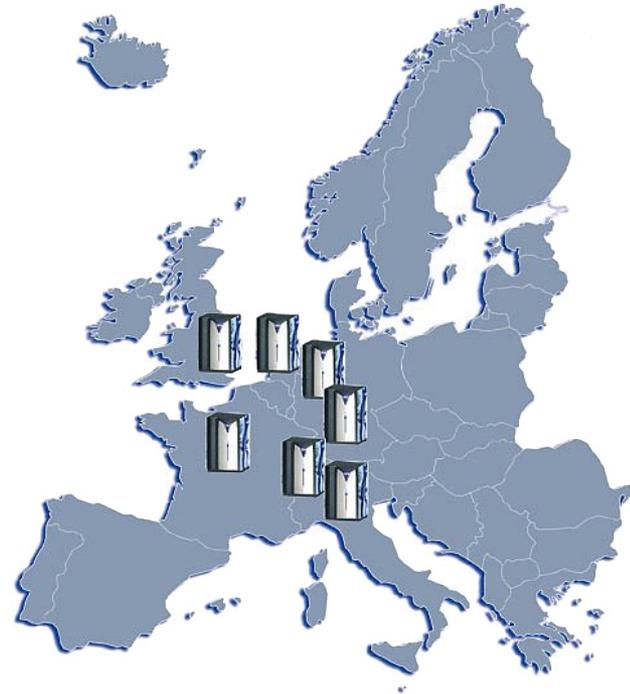
Amsterdam / The Netherlands

London / United Kingdom

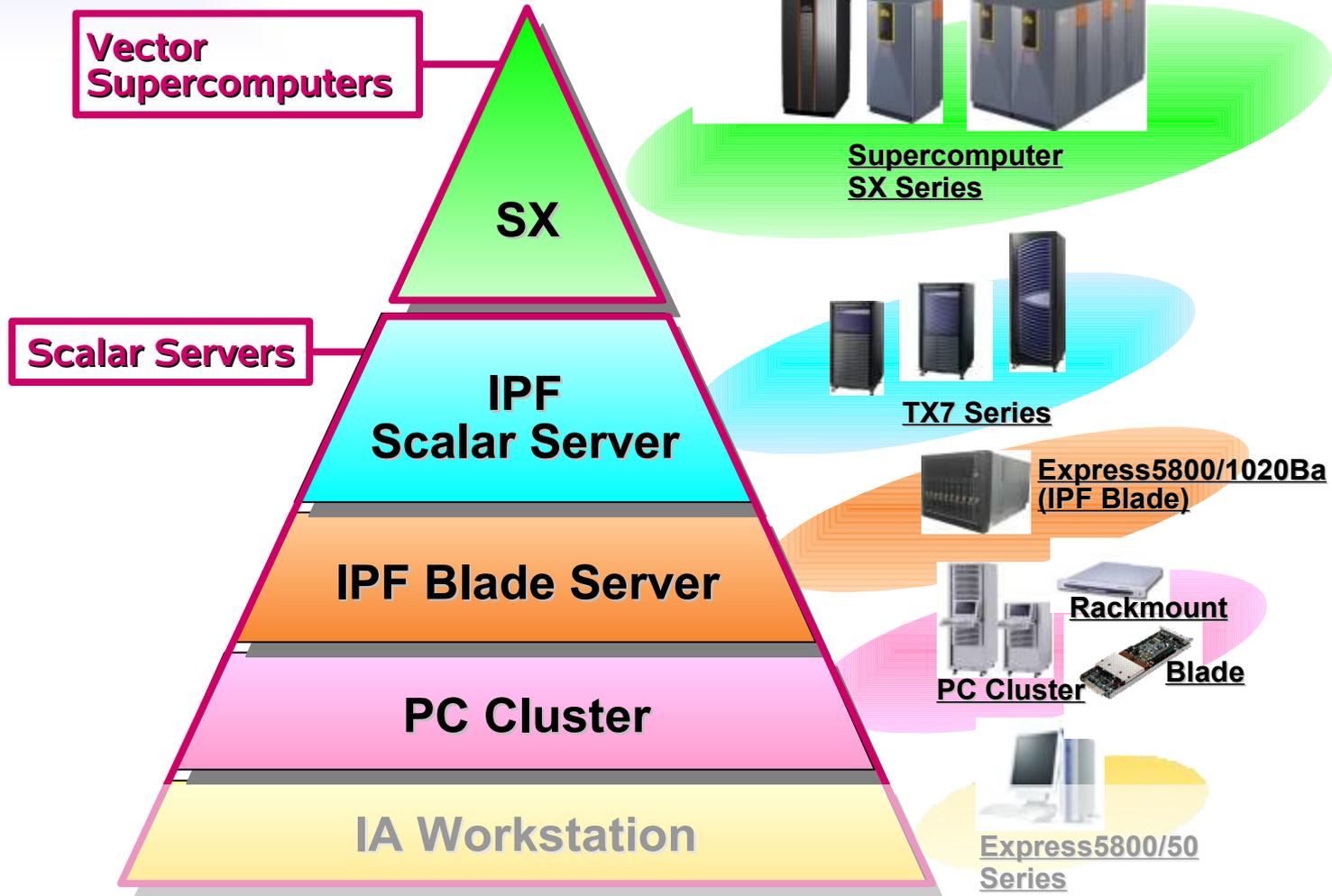
Paris / France

Lugano / Switzerland

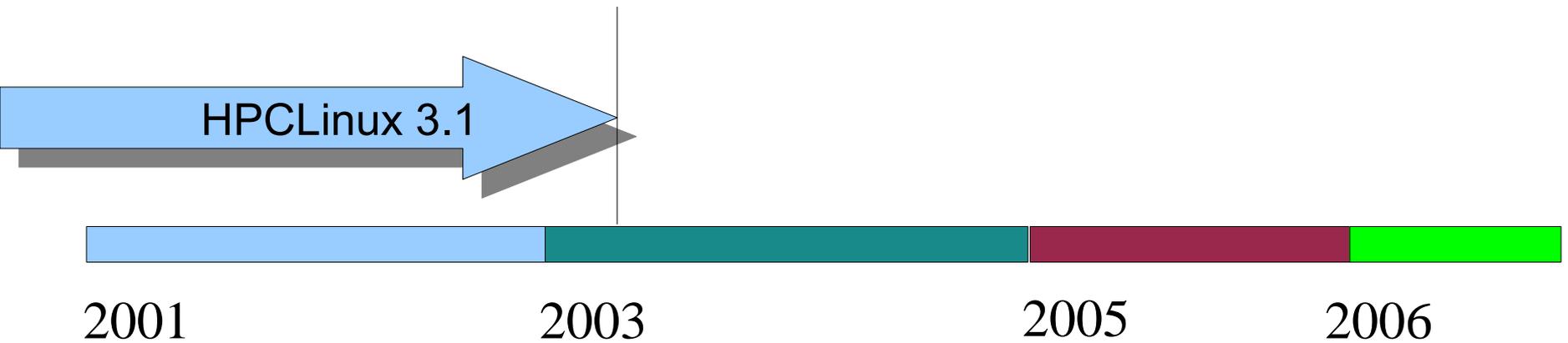
Milan / Italy



# NEC HPC Product Lineup



# OSCAR-Pro history



# OSCAR-Pro history

- RedHat based
- Kickstart installer
- manual installation of clustering packages (MPIs, C3, PBS, ...)
- HPCL3.1: automatic config

HPCLinux 3.1



# OSCAR-Pro history

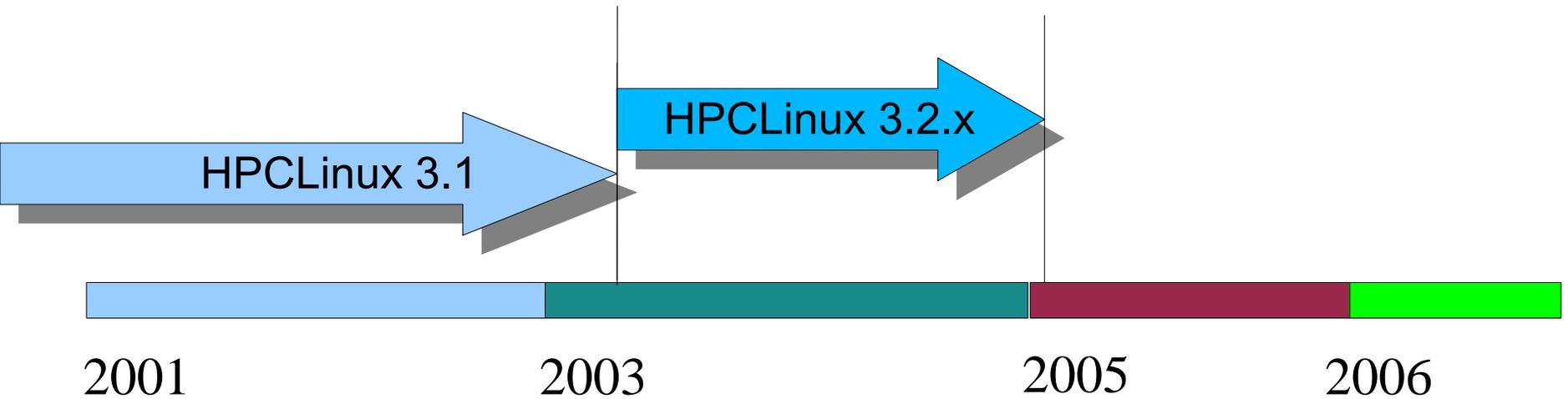
- RedHat based
- Kickstart installer
- manual installation of clustering packages (MPIs, C3, PBS, ...)
- HPCL3.1: automatic config

- reproducibility of the installation?
- quality? effort?
- Opteron coming up: only distro supporting it was SuSE9.0 (8.2)

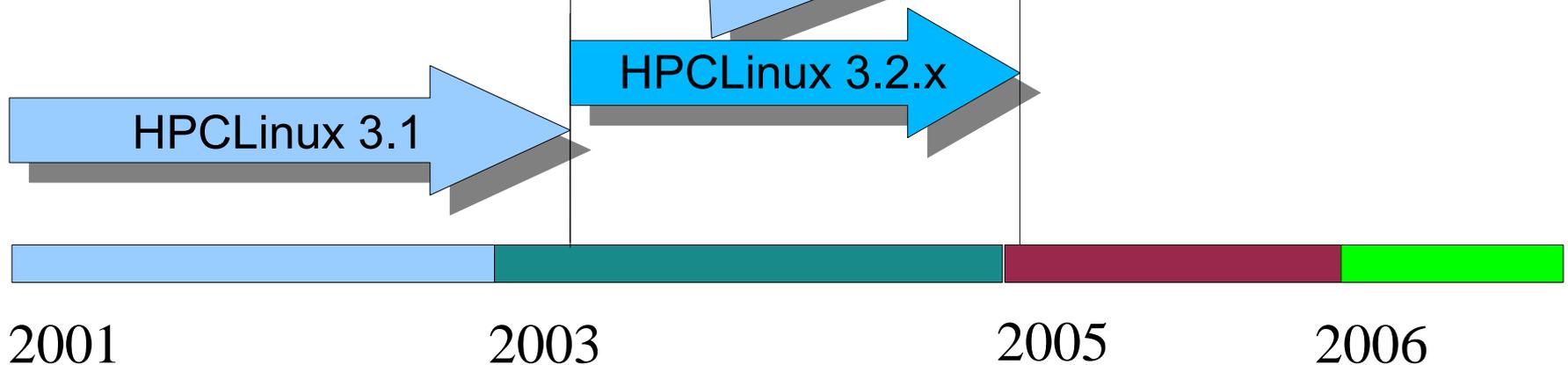
HPCLinux 3.1



# OSCAR-Pro history

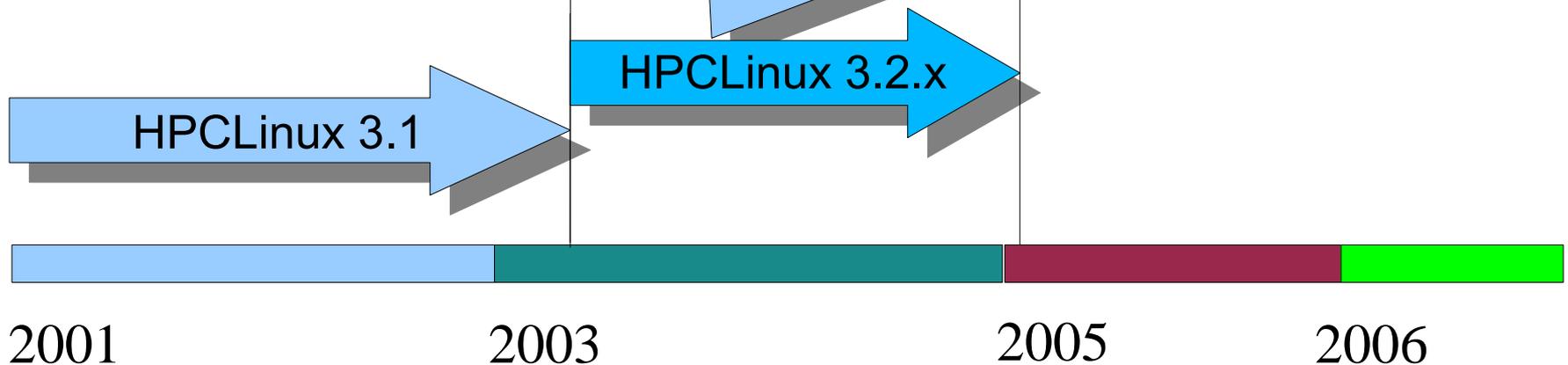


- OSCAR 2.3.1 fork
- many OSCAR 3.0 packages
- port to SuSE 9.0, SLES8, RHEL3
- x86\_64 and ia64 ports
- add-ons: Torque, Ganglia, Nagios, software RAIDs, linux-ha (!), scalability, cluster of clusters, ...
- integration



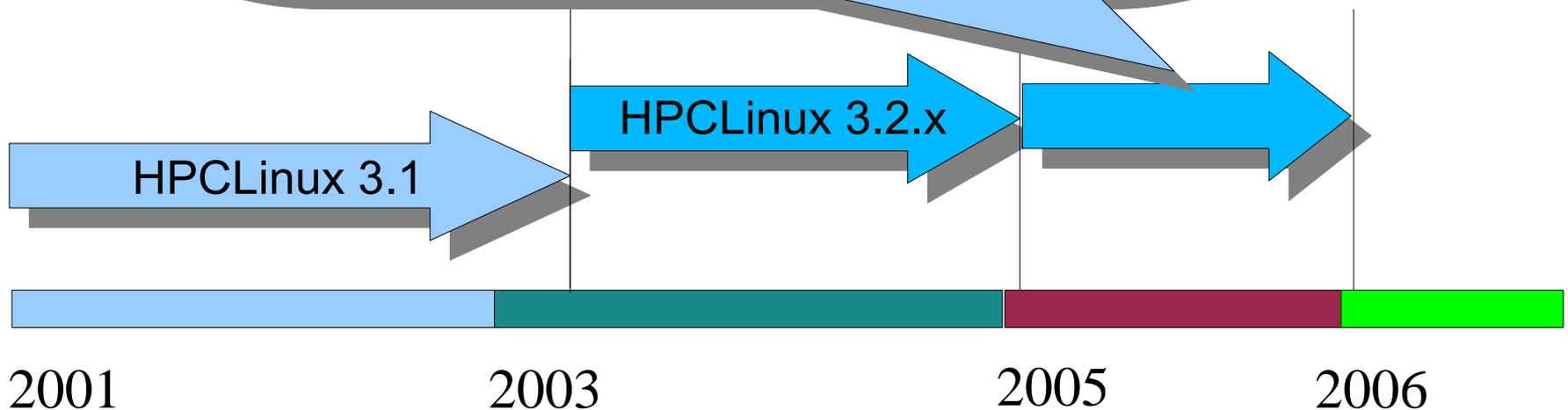
- OSCAR 2.3.1 for
- many OSCAR 3.
- port to SuSE 9.0
- x86\_64 and ia64
- add-ons: Torque software RAID, I scalability, cluster
- integration

- OSCAR 3.0, 4.0, 4.1, ...
- hard to keep up fork with OSCAR infrastructure changes
- maintenance of add-ons ties resources which are missing for R&D
- testing multitude of versions is time consuming



## Strategy change!

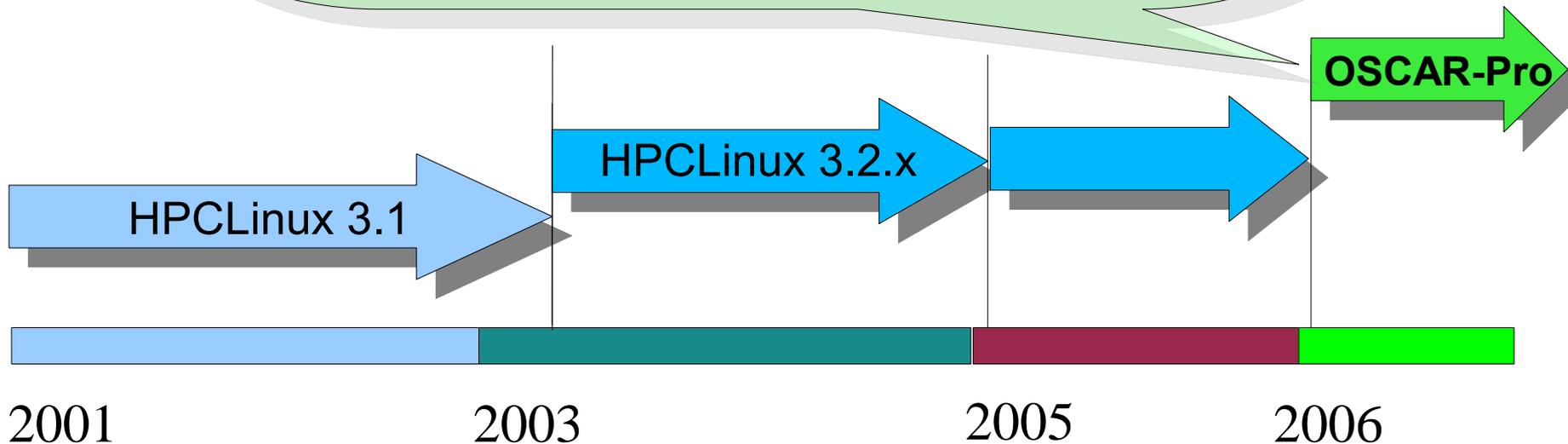
- integrate developments back into OSCAR
- NEC joined OSCAR core organizations
- contributed to OSCAR 4.2 (rel.: Nov 2005)
  - x86\_64 support, sw-raid, rhel4, packages
  - fixes
  - testing, QA



# OSCAR-Pro history

## OSCAR Pro 4

- built strictly on top of OSCAR (4.2)
- fixes and developments for OSCAR5
- more packages donated to OSCAR
- some new developments directly for OSCAR



# OSCAR-Pro Development Philosophy

- Use *best known methods* for building, installing, managing and running clusters
- Open Source! (where appropriate)
- Be open: for additional components and integration with closed source products
- **Integrate**
  - Autoconfigure components in reasonable way
  - Standardize: minimize effort once a new solution has been developed
  - ISV application ready (for industrial customers)
- Redundancy and high availability
- Support various architectures, multiple Linux distributions
- ***Flexibility, scalability and high performance!***
- ***Provide highly customized complex solutions with limited additional effort***

# Why Open Source?

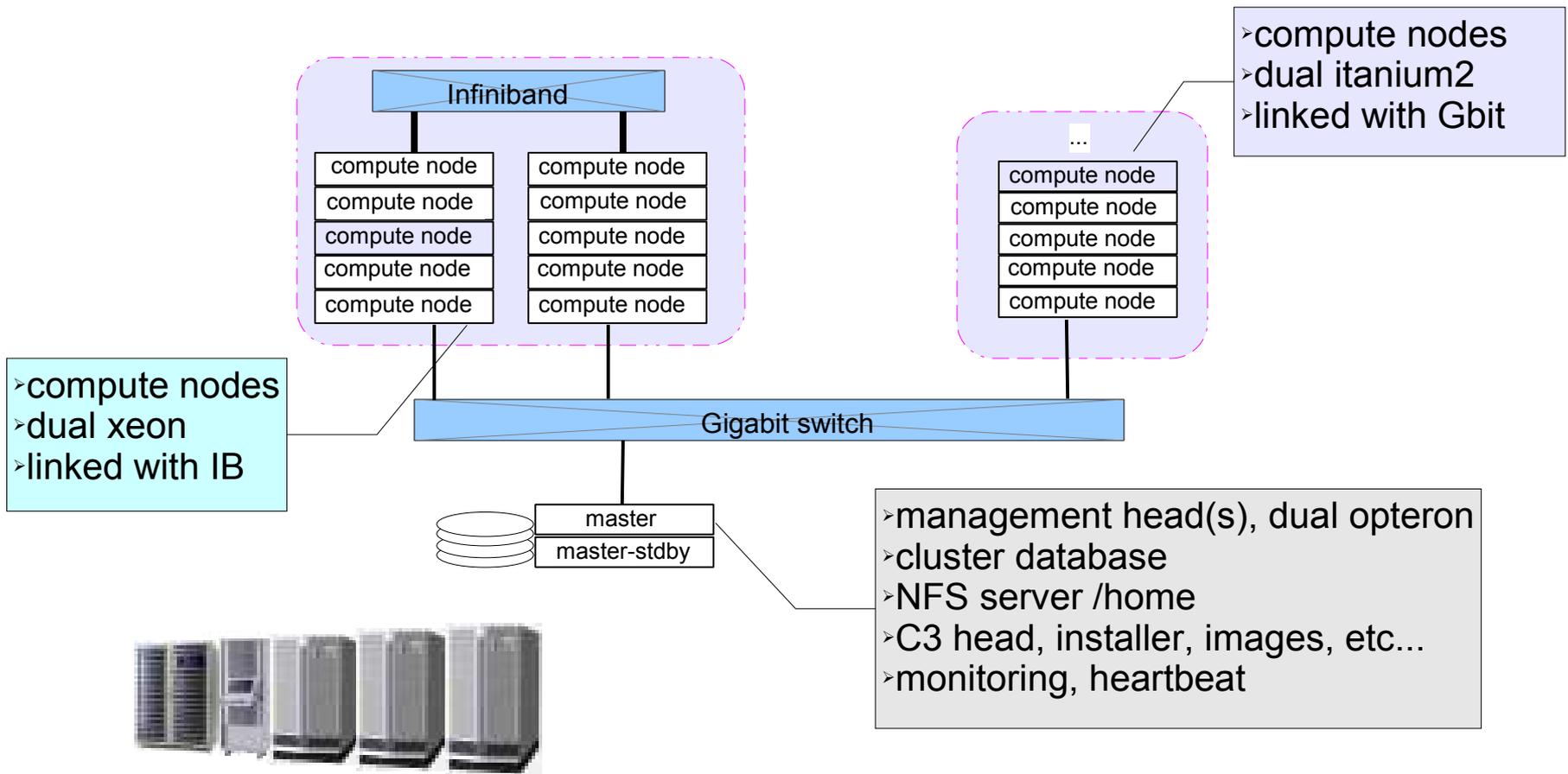
- Don't “re-invent the wheel”
  - minimize effort and development expenses
  - only expertise on the used open source solution is needed, no new design, development, team, man-power
  - faster “time to solution”
- Carefully chose solutions: proven, known
  - high quality through many users and testers
  - free marketing benefit when using well-known solutions
- Own improvements: added value for customers
  - development effort needed, but limited
  - full control over own version, but still open source
- Contribute to the open source development:
  - reduces support and maintenance costs
  - credibility for NEC open source activities
- Customer safety: full insight and control of what they are using
- Customers can choose NEC service, or do it themselves

# Customer examples

- ◆ Sometimes customers want ...
  - ◆ ... multiple architectures within the same cluster
    - ◆ Xeon, IA64, Opteron, Nocona
    - ◆ Extension of existing cluster (no additional mgmt node)
    - ◆ Safe migration environment
  - ◆ ... different distributions inside the same cluster
    - ◆ RHEL3, RHAS2.1, SUSE, FC, RH9, ...
    - ◆ e.g. because applications are validated by ISVs for different distros
  - ◆ ... multiple interconnect types in same cluster
    - ◆ Myrinet, IB, Gbit, ...

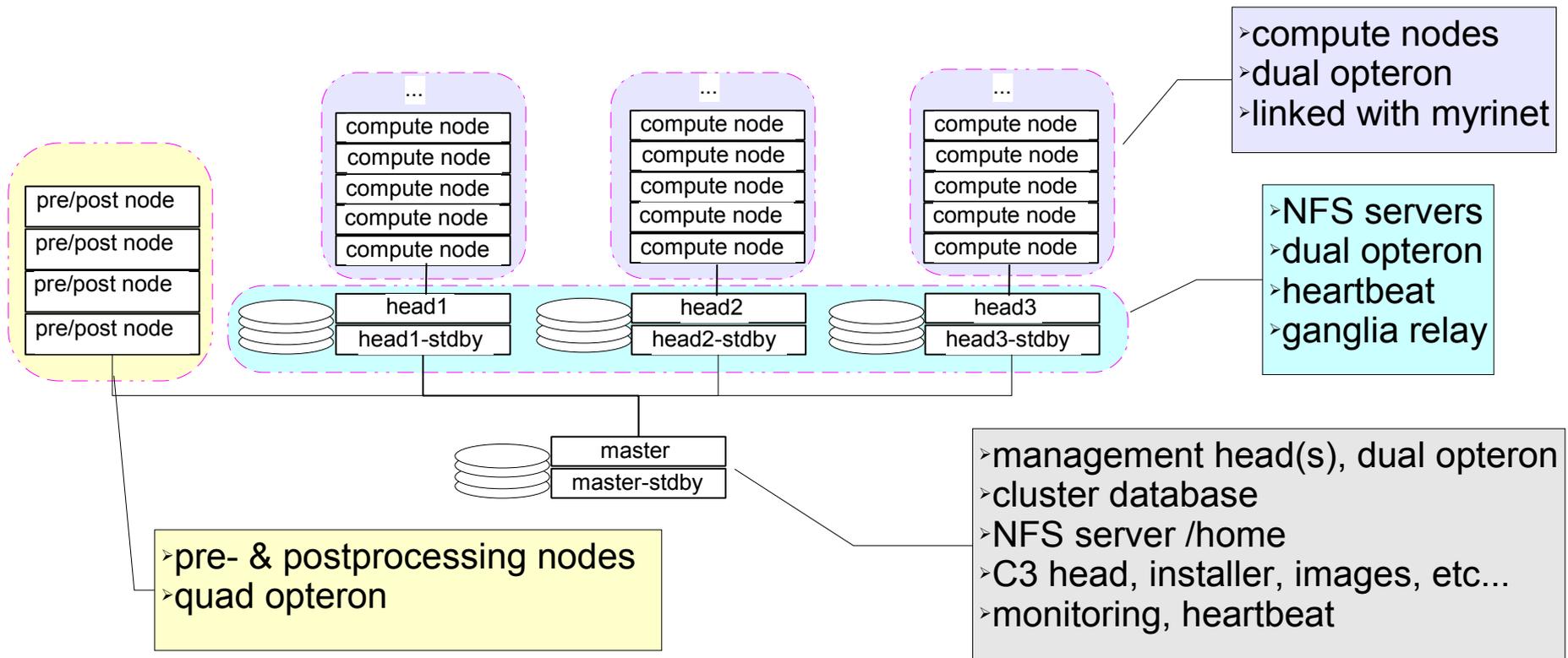
# Customer examples (1)

- ▶ Subclusters have different architecture (and distro)



# Customer examples (2)

- Subclusters have different role: running different applications



# Challenges

- ◆ Cannot charge much money for cluster software in HPC!
  - ◆ Solution must provide a cheap entry price
  - ◆ Earn money with **services**, solution must be ready for that
  - ◆ ***The Linux Way of Business***
- ◆ ! Hardware changes every 6 months ! (more or less)
- ◆ Feature richness in software is key differentiator:
  - ◆ Support various distributions
  - ◆ Flexibility in supported hardware and configurations
  - ◆ Manageability
- ◆ Satisfy the cluster administrators!

# Strengths

- **Flexibility**
  - Image based installer: different distros, different architectures, safe and controlled upgrading of client nodes
  - Heterogeneous clusters support
  - Easy to extend and integrate additional hardware and software
  - [ROCKS: tied to particular RedHat rebuild and anaconda installer]
- **Scalability**
  - Scalable deployment: atftp, multicast, (soon: bittorrent)
  - Scalable monitoring and management infrastructure
- **Reliability**
  - Mirrored disks
  - High availability (not for every customer!)
- **Ready to use**
  - When installed by trained team
- **Sound open source basis: OSCAR**

# OSCAR-Pro 4 Features (1)

- **All OSCAR 4.2 features**
- Nagios : hardware health monitoring
- Gangmet + Gangnag : Nagios – Ganglia coupling
- sensorsd : Configurable ganglia metrics feeder
- mdassist : Monitor software raids, instructions for handling failures
- Yume : high level package management tool for rpm based distros
- Sync\_files : improved version for heterogeneous setups
- SISreload : synchronize SIS database state with cluster nodes
- SC3 : (Scalable) Sub-Cluster Command and Control
- Mpicleanup : Cleanup for killed or crashed MPI zombies
- Lamrsh : Allows usage of multiple LAM MPI versions in parallel
- PBSpro : Altair's PBS pro resource manager
- NetBootManager : GUI for managing network booted nodes

# OSCAR-Pro 4 Features (2)

- Apctool : Manage power control for cluster via APC PDUs
- Cpower : Manage power control for NEC IPF Blades (through CMM)
- IPMIpower : Manage power control for clusters with IPMI BMCs
- Gscratch : global namespace through cross-automounted /scratch directories
- MPICH : 1.2.7 for gcc, pgi, intel compilers
- LAM MPI : several versions, for gcc, pgi, intel compilers
- Cluster / Image management add-ons
- High Availability (master, if required)
- Myrinet-MX/GM
- Infiniband

# Roadmap

- ◆ scalability improvements
  - ◆ 2000 nodes cluster
- ◆ heterogeneity
  - ◆ integrate SX vector supercomputers
- ◆ administration
  - ◆ CLI and GUI
- ◆ add parallel filesystem support
  - ◆ Lustre package
- ◆ Single System Image
  - ◆ Kerrighed
- ◆ from pure cluster toolkit to datacenter management

# Open Source Software: Academia & Industry Cooperation

- ◆ Industry involvement in open source development is important:
  - ◆ integrate customer requested features
  - ◆ push and drive productization
  - ◆ commit resources
  - ◆ increase credibility and popularity
  - ◆ But:
    - ◆ decisions controlled by benefit and „return of investment“
- ◆ Academia & industry complement each other
  - ◆ research & papers
  - ◆ customers & products & solutions

# Conclusions

- OSCAR is an excellent cluster infrastructure
  - feature-rich, extensible, flexible, good roadmap
  - good platform for building professional solutions
- Endorsement of OSCAR was the right decision!
- Open Source approach is important!
  - fork of OSCAR
    - steep development curve (because of full control)
    - high resource demand for maintenance
    - ties too many resources for development
  - developing with and for OSCAR
    - less maintenance effort
    - more influence on the development direction
- Marketting speak: „win-win situation“