# CLIC
# **C**luster **LI**nux pour le **C**alcul

http://clic.mandrakesoft.com

Yves Denneulin (@imag.fr)
ID-IMAG  Projet Apache
http://www-id.imag.fr

Laboratoire Informatique et Distribution
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
Institut National Polytechnique de Grenoble
INRIA INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE
GRENOBLE 1 UNIVERSITÉ JOSEPH FOURIER SCIENCES,TECHNOLOGIE,MÉDECINE

# CLIC project

Clusters + Linux : great but not easy
(management, use, programming tools)

Three partners:

- Bull : mainframe and IA64 experts
- ID : cluster experts + users
- Mandrakesoft : Linux experts

Funded by the French government (RNTL)

# Challenges

1)As good as others (OSCAR, Rocks, Score, Alinka, Scyld, etc)

- Classical tools : Installation, MPI, PVM, PBS, Ganglia

2) Mandrake spirit (as many packages as possible)

- Math libraries (blas, scalapack, PETSC)
- Cluster middleware (CORBA, java)
- Virtual Reality software, etc

# Challenges

3) Compatibility

- Hardware support (IA64, Myrinet, SCI, Opteron, etc) and software support (Mosix, HA, LVS, etc)

4) Scalability

- Installation (ka-deploy), parallel commands (ka-run, etc), file systems (NFSp, PVFS)

# Scalability

- How to avoid bottlenecks in large clusters
  - Cluster installation
  - Parallel commands
  - Data staging
  - File systems

# Ka-tools

- Ka-deploy
  - Deploys a cluster cloning the golden node
  - allows heterogeneous hardware configurations

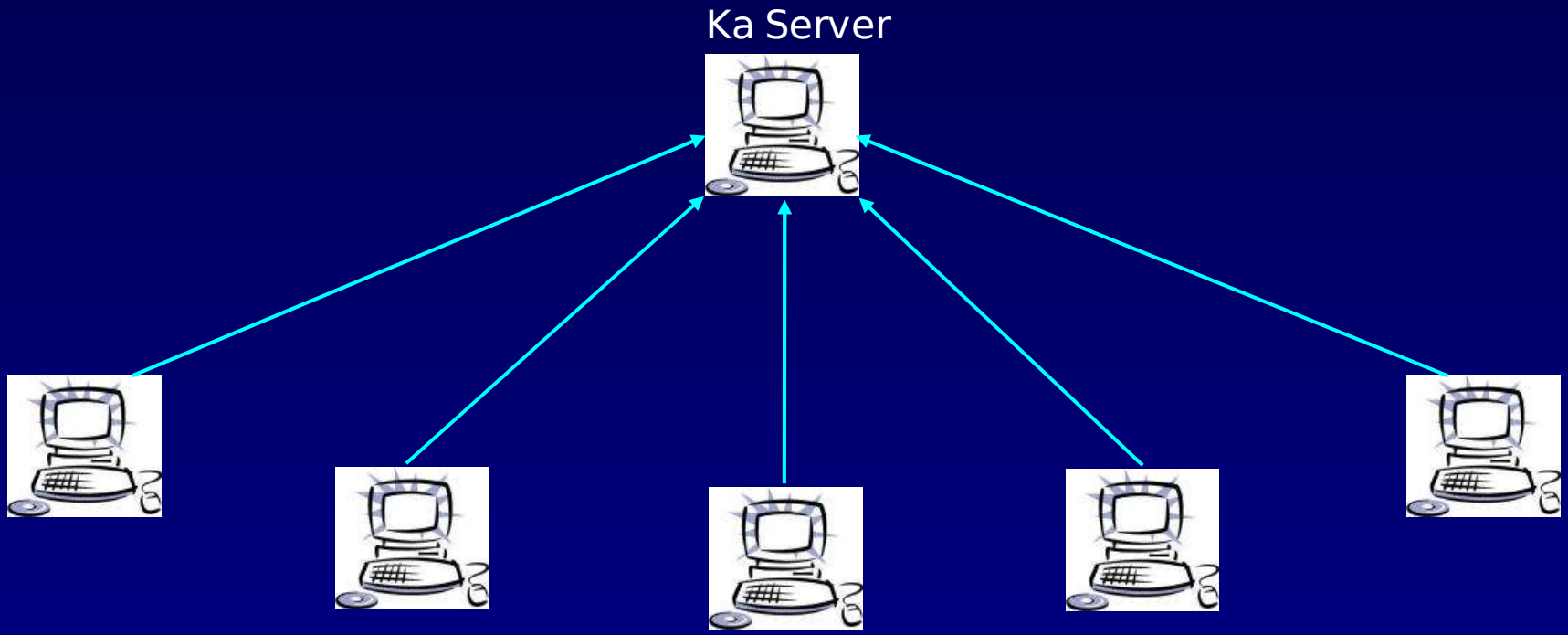
  - Ka-run
  - Parallel and distributed command launcher (rshp)
  - Optimized file copy on cluster (mput)

# Launching Ka replication

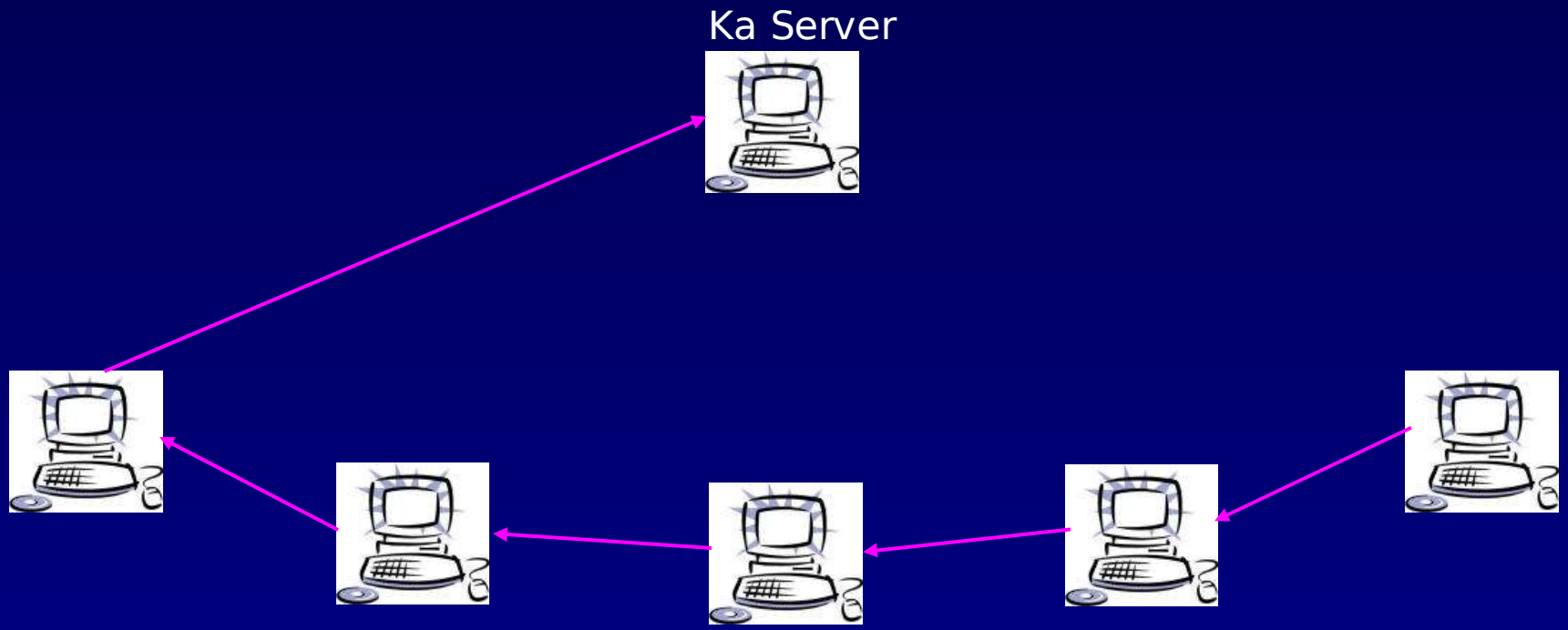
- Ka-deploy requires network boot to be used efficiently
  - a PXE/DHCP server with a tftp server
  - Network boot capable nodes

- After the boot sequence, the replication can start...

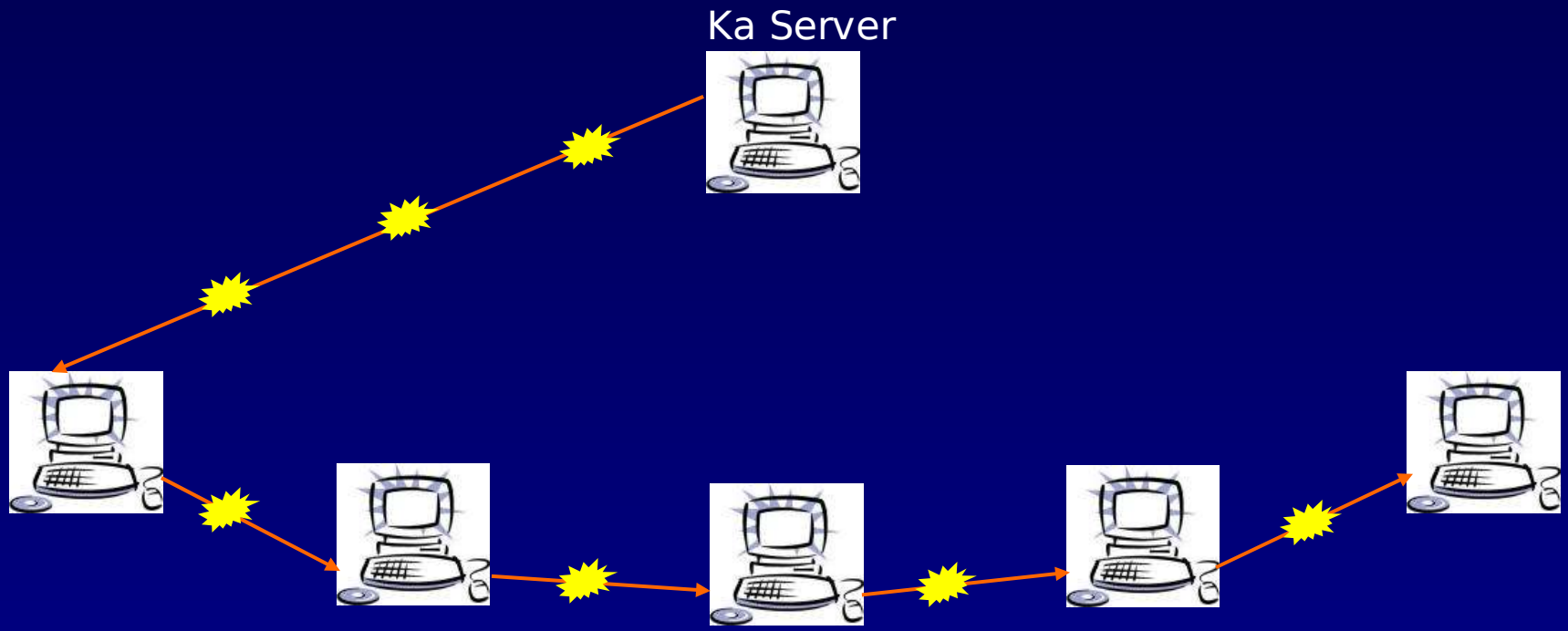
# Ka-deploy



# Ka-deploy (2)



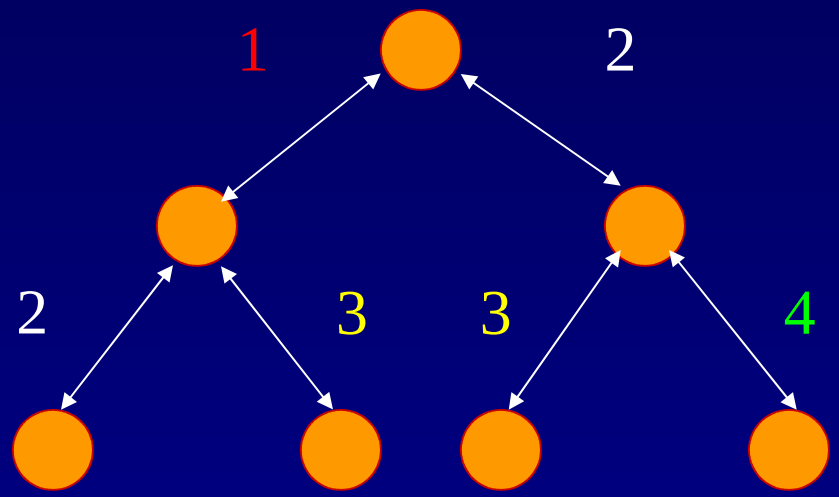
# Ka-deploy (3)



# Rshp

- What?
  - Launching one command from a server onto n remote nodes
- How?
  - Launchs distributed ssh or rsh connections
    - distribute the load betwenn the nodes
      - trees,  work stealing, etc
  - Executes the command and redirect outputs

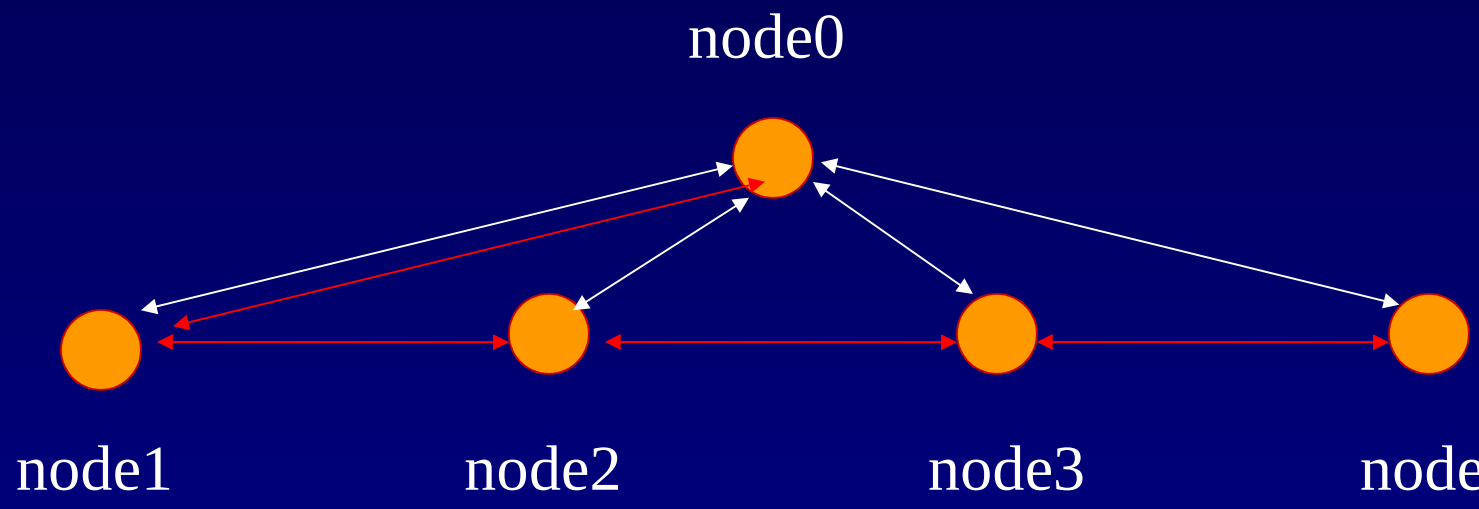
# Building the launching tree with rshp



# Mput

- Quick diffusion of a file over the cluster
- Fastest topology to connect all the nodes : flat tree
- Most efficient topology to propagate a file : chain
- Mput = mixed method
  - mput is launched by a flat tree
  - The nodes get the file through a chain

# Data staging




mput file copy command
Better than reliable multicast, trees, etc

# A cluster oriented Linux Distro.

- CLIC is a real Linux Distro
  - Glibc, kernel, ...
- But it also includes
  - Auto-configuration tools (clusterscripts)
  - Parallel tools (rshp, mput, urpmi parallel)
  - MPICH, LAM/MPI, PVM
  - Automated duplication tools (KA)

# Parallel Urpmi

- Use rshp, mput and urpmi to quickly update a cluster to maintain the same set of rpms over the cluster

- rshp and mput use case
  - Rshp launch urpmi on all the nodes to compute the dependancies

# Parallel Urpmi (2)

- rshp and mput use case
  - All these dependancies are analyzed to get the maximum set of rpms to mput over the cluster
  - Each node independently installs all the dependencies it needs to install the target rpm

  - The same tools exists to uninstall rpms over the cluster

# CLIC Status

- Phase  2 is about to be released
- New backend fully written in Perl
- New installation procedure (Just Power On)
- Dual ethernet cluster
  - Separated computing & administration network
- New urpmi/urpme parallel

# Annexes

# Parallel vs distribute

| | Rshp client asynchrone | Rshp 200 threads | Rshp binomial | Rshp binaire | Rshp (ssh + processus) arité 4 | C3 rsh | C3 ssh |
|---|---|---|---|---|---|---|---|
| **Tps sec** | 0.45 | 0.49 | 1.4 | 1.6 | 2.19 | 2.57 | 14.52 |
| **Travail (%CPU x Tps)** | 27.9 | 32.2 | <7 | <8 | 10.3 | 243.9 | 1057 |
| **Nb processus** | 1 | 1 | 1 | 1 | 5 | 201 | 201 |

RSHP «  null command  » on 200 nodes

# Mput performances 200 nodes

| | Mput 1Mo | Mput 2Mo | Mput 4Mo | Mput 8Mo | Mput 16Mo |
|---|---|---|---|---|---|
| Tps sec | 0.97 | 1,09 | 1.35 | 1.81 | 2.82 |

Bandwidth :

- Mput 200 nodes: 8 MB/sec (latency 0,8 secs)
- point to point network: 12 MB/sec
- Limited by disk bandwith

# File systems in CLIC

- Direct Attached Storage aggregation
- One meta data server, several data servers : PVFS, NFSp (NFS compliant)