Kerrighed : un système d'exploitation SSI pour grappes

Christine Morin
Equipe-projet PARIS
INRIA Rennes - Bretagne Atlantique

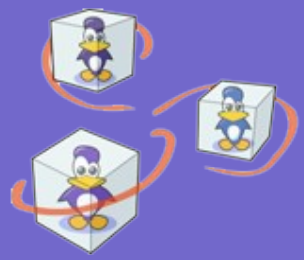
20/10/2008 – ENVOL -- Annecy

40 ans la révolution de l'information

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

INRIA

PARIS Project-Team



Kerrighed

Linux clusters made easy

Plan

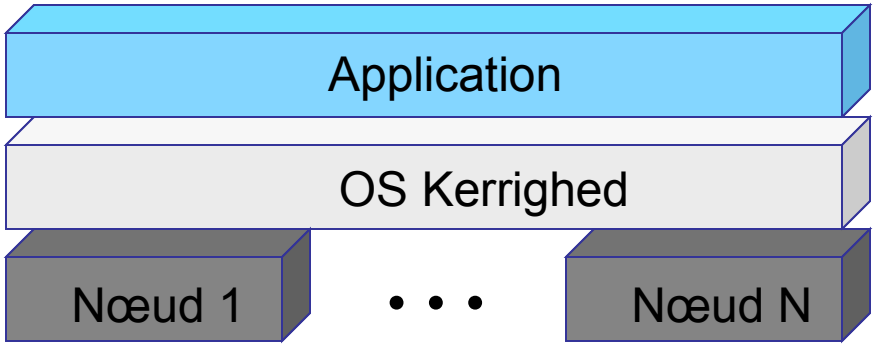
- Généralités
- Gestion globale des processus
- Gestion globale de mémoire
- Gestion globale des fichiers
- Gestion globale des communications
- Paramétrage du système
- Administration de la grappe
- Gestion des défaillances
- Expérimentations
- Conclusion

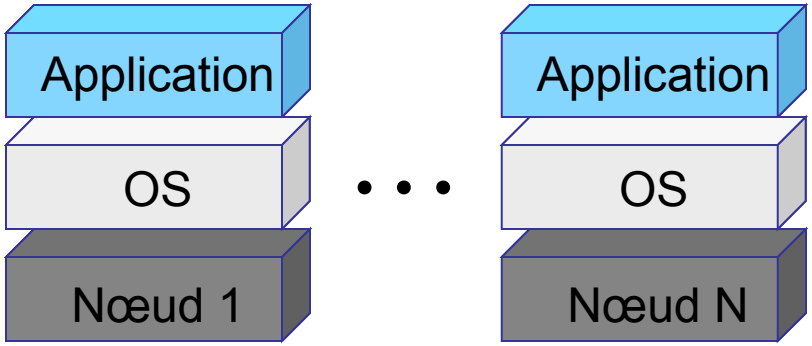
Le système Kerrighed (1/3)

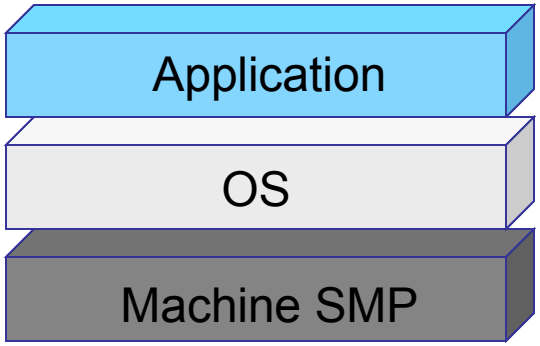
- Système d'exploitation pour grappes de calculateurs

Le système Kerrighed (2/3)

- Système à image unique
  - Illusion d'une machine multiprocesseur à mémoire partagée
    - SMP virtuel
  - Exécution d'applications patrimoniales sans modification ni recompilation
  - Interface système standard

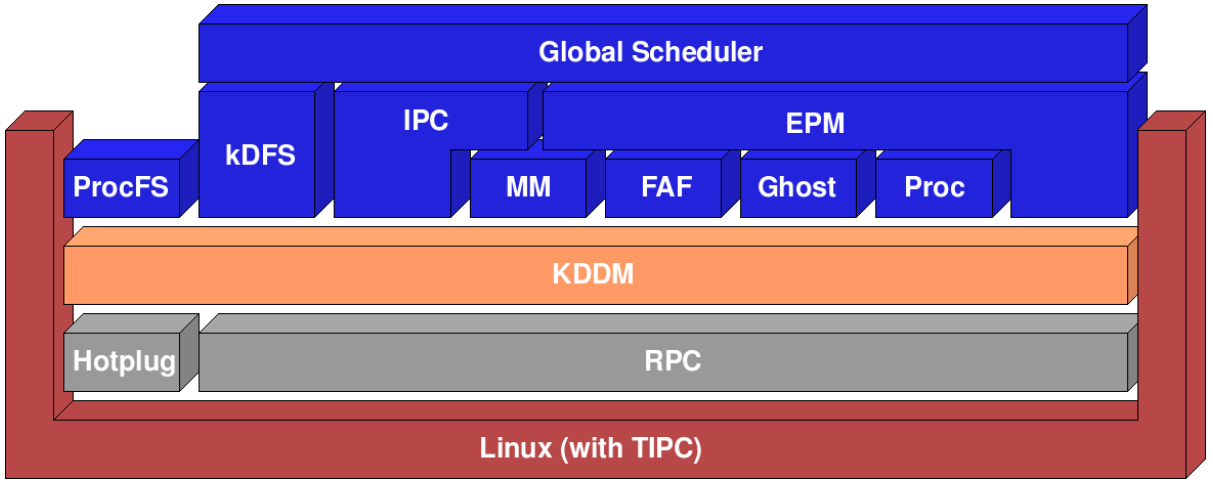




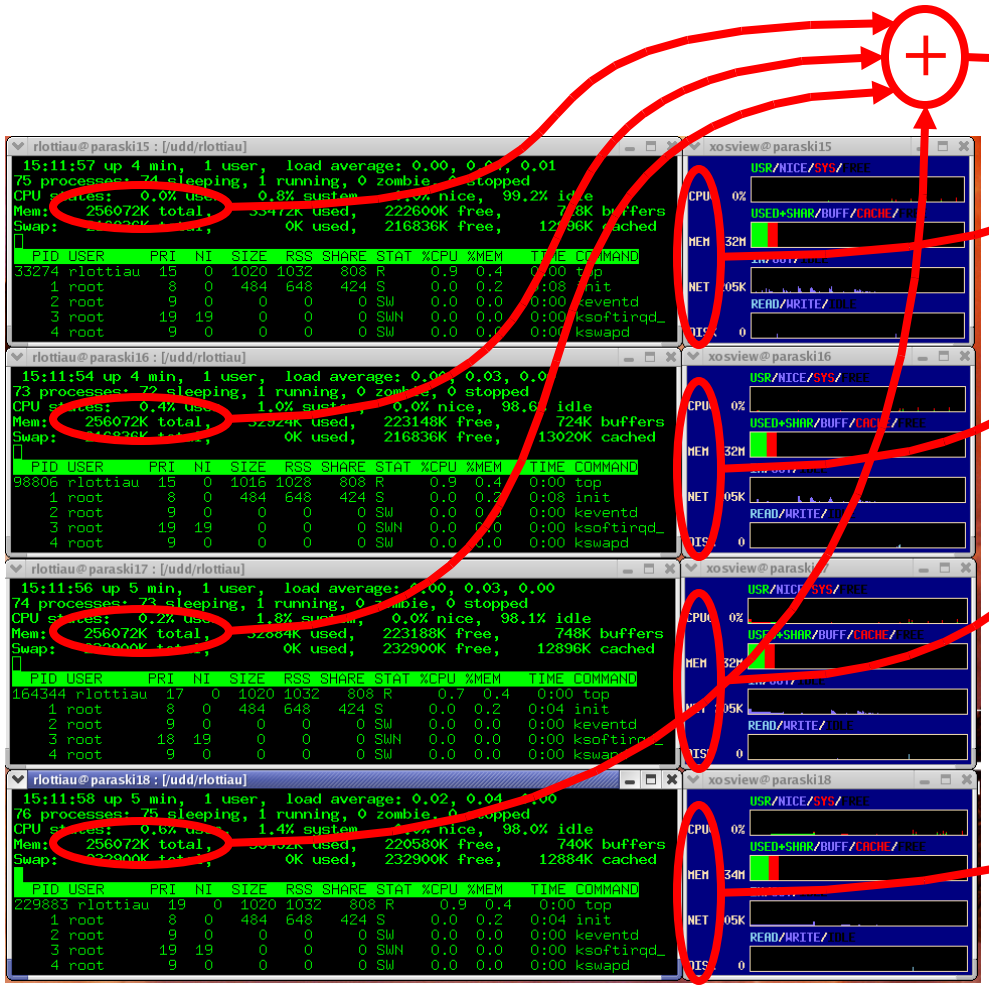


Le système Kerrighed (3/3)

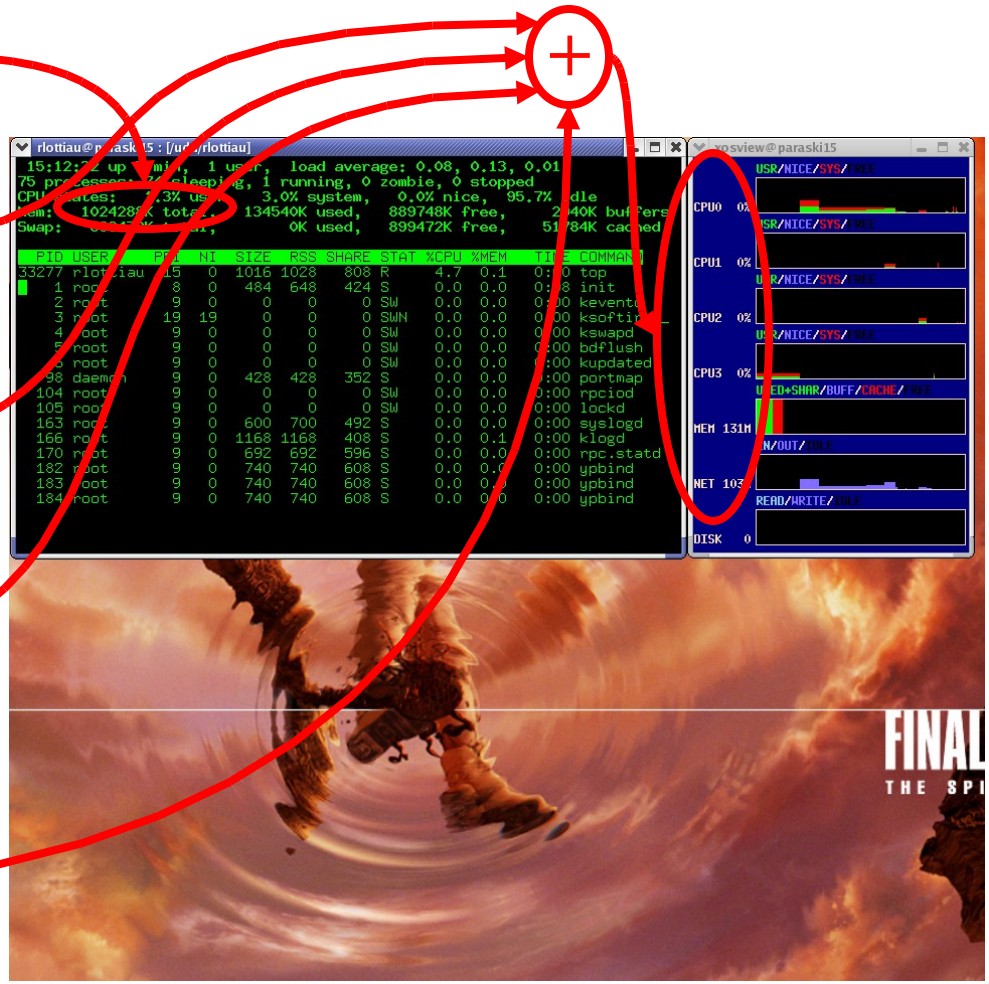
- Mise en œuvre
  - Ensemble de services distribués pour la gestion globale des ressources
  - Implémentation par des modules noyau ajoutés au système Linux et modification du noyau Linux
  - Interface Posix



Linux standard vs Linux/Kerrighed



Linux standard



Linux/Kerrighed

Utilisation de l'outil XOSVIEW **non modifié**

Autres systèmes SSI fondés sur Linux

- Mosix (mosix.org)
  - Hebrew University, Jerusalem
  - Prof. Amnon Barack
- openMosix (openmosix.sourceforge.net)
  - Issu du projet Mosix
  - Moshe Bar
- OpenSSI (openssi.org)
  - Bruce Walker, HP

Plan

- Généralités
- Gestion globale des processus
- Gestion globale de mémoire
- Gestion globale des fichiers
- Gestion globale des communications
- Paramétrage du système
- Administration de la grappe
- Gestion des défaillances
- Expérimentations
- Conclusion

Espace global de processus

- Identification unique des processus à l'échelle de la grappe
- Commandes traditionnelles à l'échelle de la grappe
  - kill
  - ps
  - top
    - Disponibles sur tous les nœuds de la grappe
    - Agissent sur tous les processus de la grappe quelque soit leur localisation
- Opérations sur les processus
  - Création locale ou à distance
  - Migration
  - Sauvegarde et restauration de points de reprise
  - Clonage

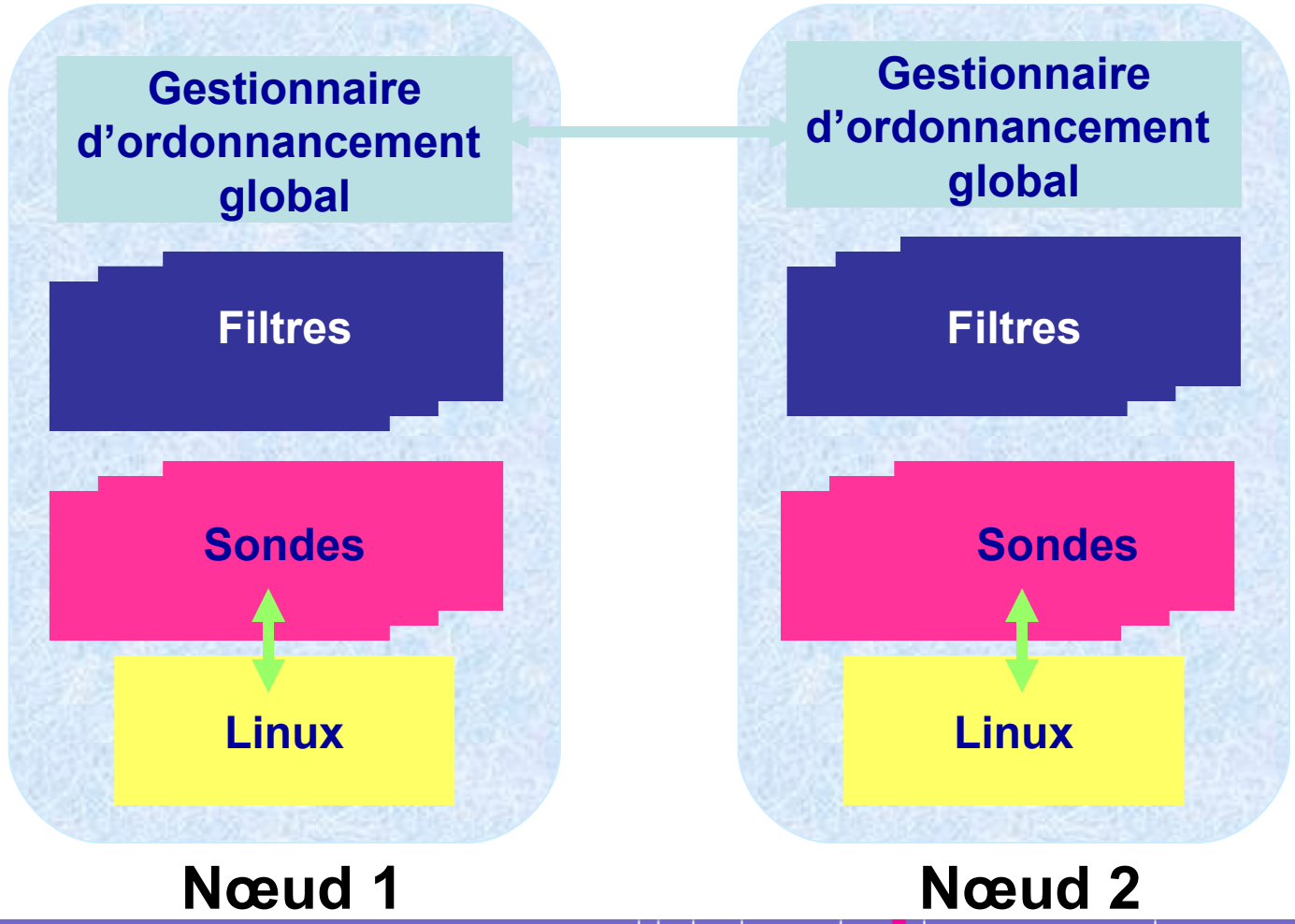
Ordonnancement global de processus

- Placement de processus
  - Choix du nœud d'exécution d'un processus lors de sa création
- Equilibrage et partage de charge
  - Répartir les processus entre les nœuds
  - Déplacement de processus en cours d'exécution
- Ajout ou arrêt d'un nœud pour maintenance ou gestion de l'énergie
  - Déplacement des processus locaux en cours d'exécution avant l'arrêt d'un nœud
  - Déplacement de processus sur le nœud nouvellement intégré à la grappe
- Politique d'ordonnancement
  - Quels critères prendre en compte pour évaluer la charge d'un nœud ?
  - Quand déplacer un processus ?
  - Quel processus déplacer ?
  - Vers quel nœud déplacer le processus choisi ?

Ordonnanceur global configurable

- Spécialisation de la politique d'ordonnancement
  - Utilisation de l'interface configFS
    - Interface système de fichiers pour créer, configurer, détruire des objets noyau
      - mkdir, cat, echo, read, write
- Politique d'ordonnancement
  - Sondes, filtres, gestionnaire d'ordonnancement global
- Changement à chaud de la politique d'ordonnancement global
  - Sans arrêter la grappe
  - Sans arrêter les applications en cours d'exécution

Ordonnanceur global modulaire



Plan

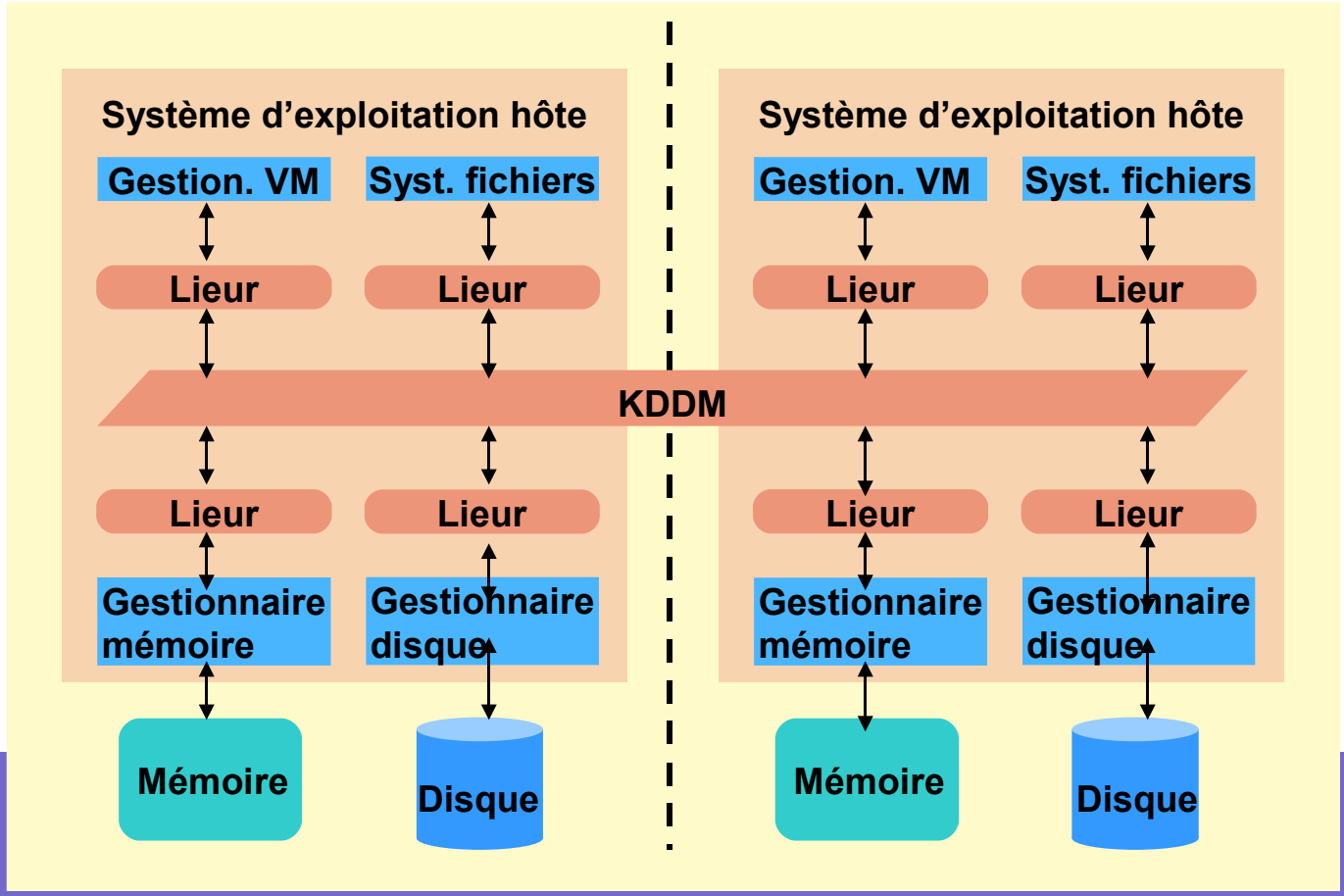
- Généralités
- Gestion globale des processus
- Gestion globale de mémoire
- Gestion globale des fichiers
- Gestion globale des communications
- Paramétrage du système
- Administration de la grappe
- Gestion des défaillances
- Expérimentations
- Conclusion

Gestion globale de la mémoire : fonctionnalités

- Partage de mémoire entre des processus ou threads s'exécutant sur différents nœuds de la grappe
- Aggrégation de mémoire
  - Possibilité pour un processus d'utiliser toute la mémoire disponible sur une grappe
  - Pagination en mémoire distante plutôt que sur le disque local
  - Politique de migration des procesus près des données qu'ils utilisent
- Migration de processus
  - Transfert de l'espace d'adressage du processus déplacé
- Caches de fichiers coopératifs
  - Éviter de lire sur disque un bloc de données qui se trouve déjà dans le cache de fichiers d'un nœud de la grappe

KDDM : un concept clé de Kerrighed

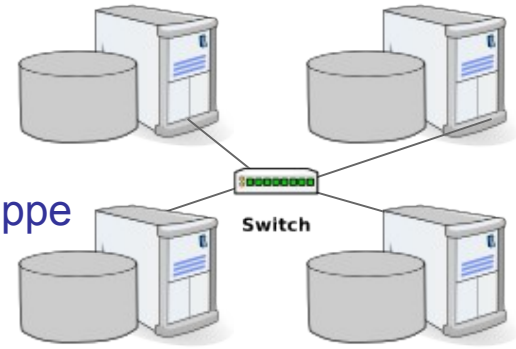
- KDDM (conteneur)
  - Accès à distance et partage cohérent de structures de données du noyau au sein de la grappe



Plan

- Généralités
- Gestion globale des processus
- Gestion globale de mémoire
- Gestion globale des fichiers
- Gestion globale des communications
- Paramétrage du système
- Administration de la grappe
- Gestion des défaillances
- Expérimentations
- Conclusion

Système de fichiers distribué parallèle - kDFS

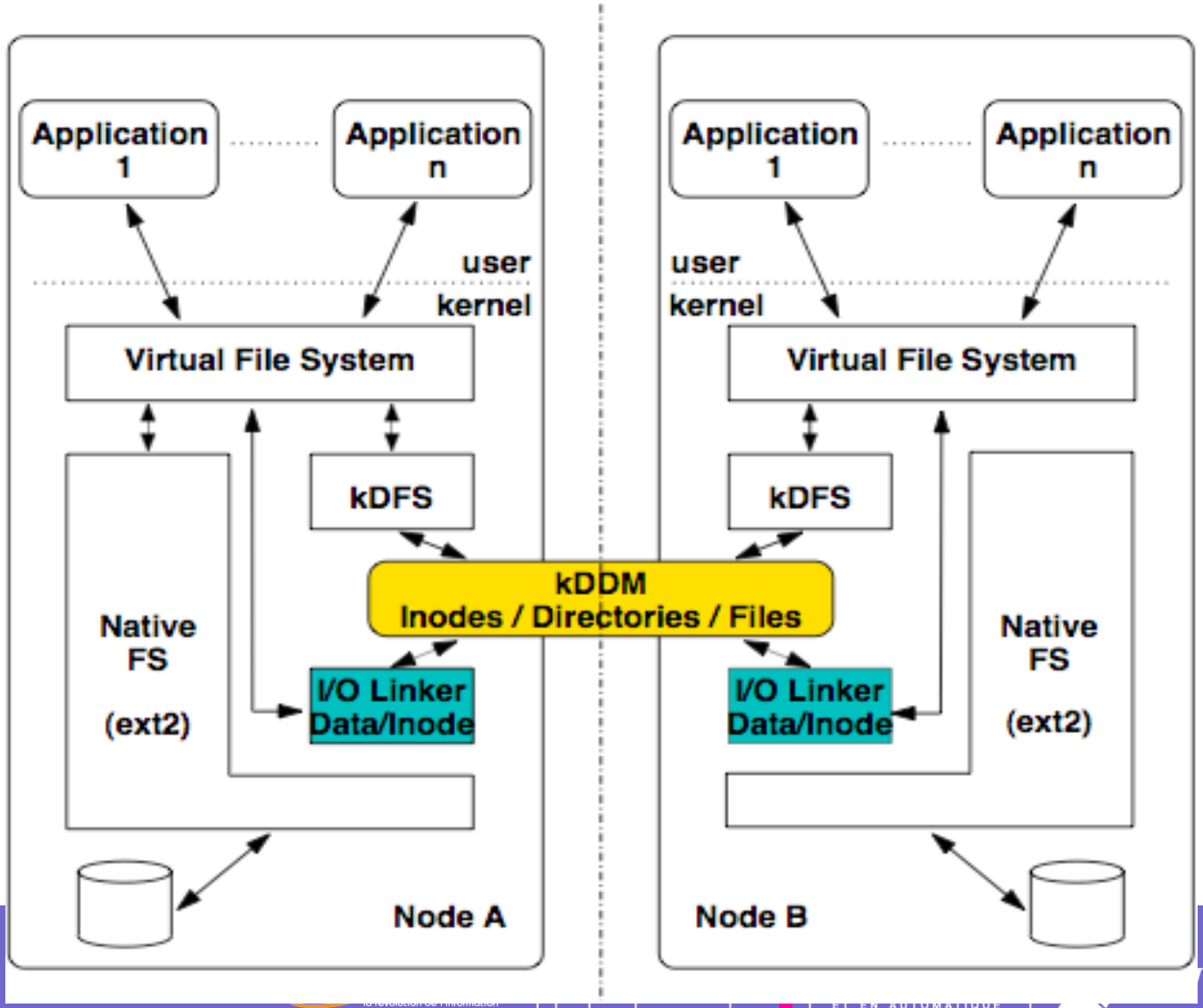


- Objectifs
  - Fédérer l'espace de stockage disponible sur les nœuds de la grappe
    - Espace de stockage unique virtuel
    - Une seule arborescence
  - Entrées/sorties disque efficaces pour les applications (parallèles) exécutées sur la grappe
    - Données des gros fichiers stockées sur des disques de différents nœuds
  - Redondance de données
  - Coopération avec les autres services du système
    - Support pour la sauvegarde efficace des données de fichiers lors de points de reprise d'applications
    - Ordonnancement global avec prise en compte du schéma d'accès des applications aux données stockées sur disque

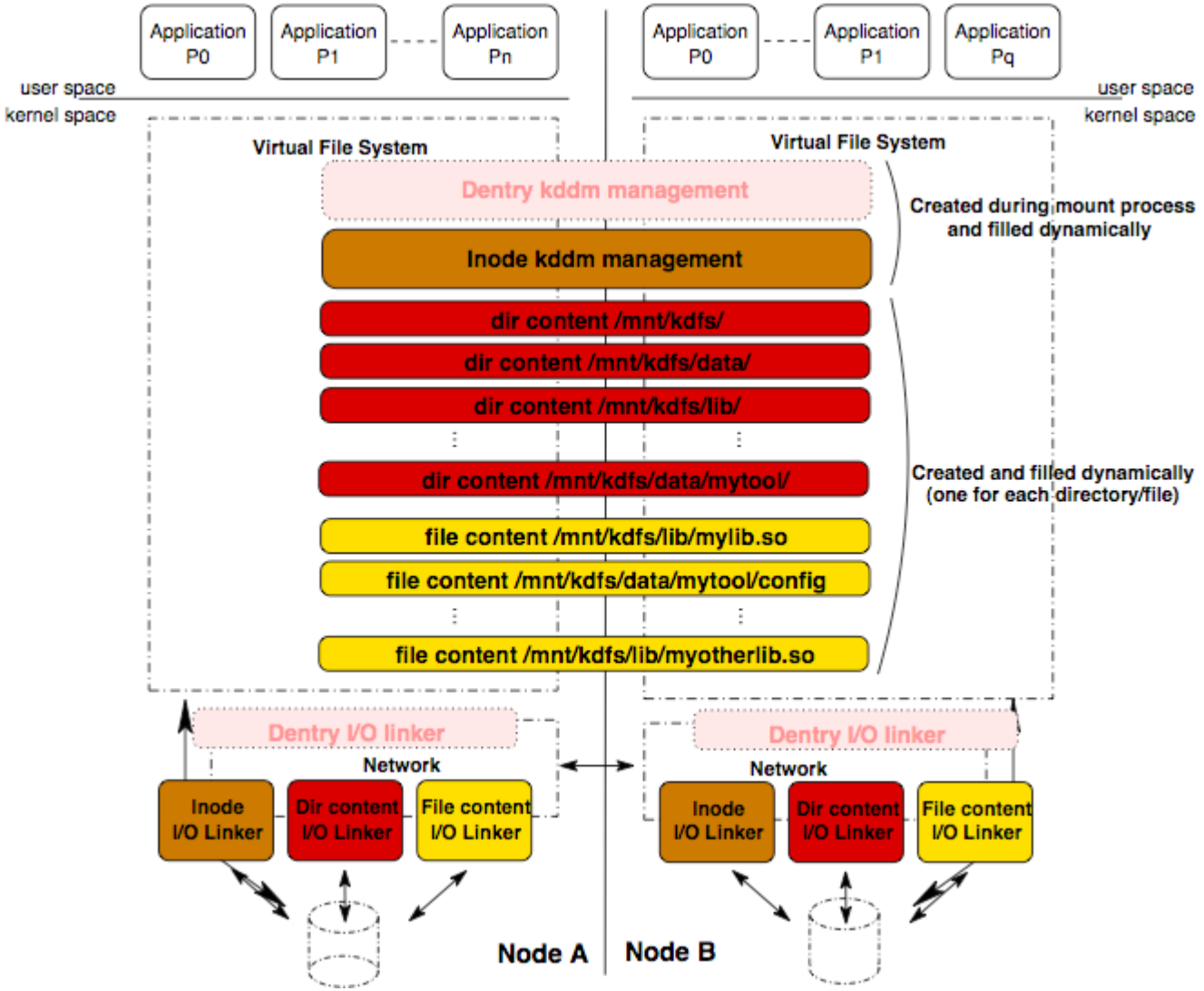
kDFS

- Principes de conception
  - Système de fichiers complètement symétrique
  - Entièrement construit au-dessus du concept de KDDM
    - N'utilise aucun autre mode de communication entre les nœuds
  - Le stockage sur disque est géré par un système de fichier local natif du système Linux
    - Ext2, ext3, …

kDFS fondé sur le concept de KDDM



kDFS fondé sur les KDDM



Plan

- Généralités
- Gestion globale des processus
- Gestion globale de mémoire
- Gestion globale des fichiers
- Gestion globale des communications
- Paramétrage du système
- Administration de la grappe
- Gestion des défaillances
- Expérimentations
- Conclusion

Gestion globale des communications

- Applications communicantes
  - Préserver les meilleures performances possibles en présence de migration de processus
    - Mémoire partagée pour deux processus sur le même nœud
    - Pas d'indirection pour deux processus sur des nœuds différents
  - Support des interfaces traditionnelles
    - Sockets, pipes, fifo, ….
- Services distribués de Kerrighed
  - Interface noyau
  - Efficacité, réactivité
    - Messages actifs
  - TIPC dans la version courante
- Support de technologies d'interconnexion variées
  - Capacité à utiliser les pilotes de cartes réseau optimisés pour les réseaux à haute performance
    - Ethernet 10G, Infiniband, Myrinet, …
  - Capacité à exécuter des applications utilisant un système de communication court-circuitant le système d'exploitation

Plan

- Généralités
- Gestion globale des processus
- Gestion globale de mémoire
- Gestion globale des fichiers
- Gestion globale des communications
- Paramétrage du système
- Administration de la grappe
- Gestion des défaillances
- Expérimentations
- Conclusion

Paramétrage du système

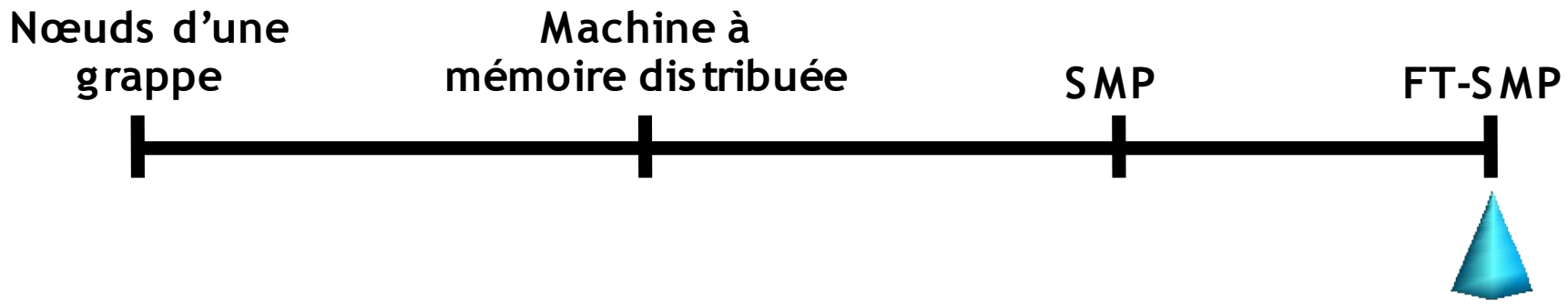
- Kerrighed offre de nombreuses fonctionnalités

- Toutes les fonctionnalités ne sont pas forcément toujours utiles
  - Toutes les applications n'ont pas besoin de tolérance aux fautes
  - Dans certains cas, il est préférable que les processus d'un groupe donné soient tous sur le même nœud
  - ...

- Activer certaines fonctionnalités peut même conduire à des dysfonctionnements !
  - Répartir sur les nœuds d'une grappe les *threads* d'une application qui n'est pas prévue pour cela
  - Sauvegarder un point de reprise pour un processus de très courte durée
  - ...

Capacités : paramètrage du SSI pour l'adapter aux besoins

- Capacités Kerrighed
  - Sélection des fonctionnalités voulues
  - Granularité : processus L'utilisation de capacités autorisées permet du contrôle et de la sécurité
- Capacités et sécurité
  - Définition par l'administrateur système de capacités par défaut pour chaque utilisateur

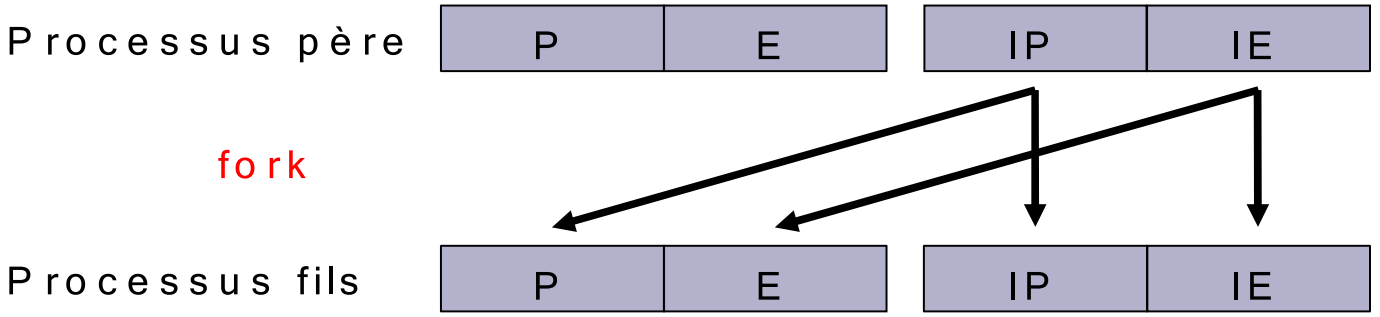


Capacités

- Deux types de capacité
  - Autorisée
  - Effective
- Pour chaque type, 2 ensembles de capacité
  - Capacités du processus courant
  - Capacités des processus fils
- Transmission des capacités par héritage

Capacités

- Chaque processus possède quatre ensembles de capacités
  - Ses capacités autorisées (P)
  - Ses capacités effectives (E)
  - Les capacités autorisées par défaut transmises à ses processus fils (IP)
  - Les capacités effectives par défaut transmises à ses processus fils (IE)



Quelques capacités

- CAN_MIGRATE
  - Définit si les processus sont autorisés à migrer
- DISTANT_FORK
  - Permet à un processus de créer des processus fils sur les nœuds distants
- CAN_CHECKPOINT
  - Permet à un processus de sauvegarder son état (point de reprise)
- SEE_LOCAL_PROC_STAT
  - Définit si un processus peut voir l'état local des nœuds
- ...

Interface

- Modification des capacités d'un processus par une commande *shell*
  - *krg_capset [pid] <options>*
- *Options*
  - *-l : affichage des capacités d'un processus*
  - *-e : Définition des capacités effectives d'un processus (E)*
  - *-p : Définition des capacités autorisées d'un processus (P)*
  - *-d : Définition des capacités effectives héritées par défaut par les processus fils du processus (IE)*
  - *-i : Définition des capacités effectives héritées par défaut par les processus fils du processus (IP)*

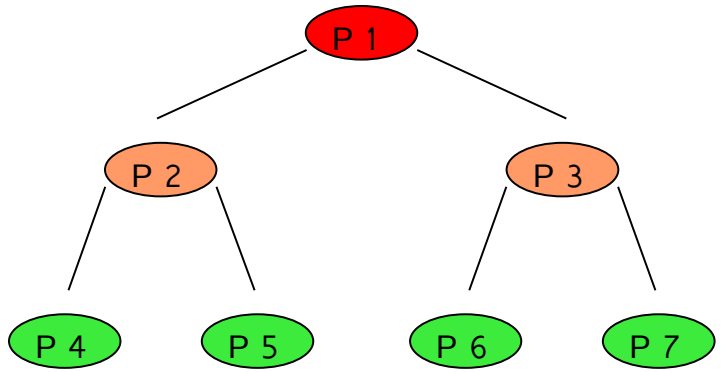
Capacités : un exemple (1)

```
paraski33% krg_capset -l
  Effective Capabilities: 043
    CHANGE_KERRIGHED_CAP, USE_CONTAINERS
  Effective Capabilities: 043
    CHANGE_KERRIGHED_CAP, USE_CONTAINERS


paraski33% cat /proc/cpuinfo
  processor       : 0
  vendor_id       : GenuineIntel
  cpu family      : 15
  ...
  processor       : 1
  vendor_id       : GenuineIntel
  cpu family      : 15
  ...
paraski33% cat /proc/meminfo
   MemTotal:      1032644 kB
   MemFree:        831052 kB
   ...
paraski33% cat /proc/meminfo
   MemTotal:      1032644 kB
   MemFree:        831052 kB
   ...
```

```
paraski33% krg_capset -d +SEE_LOCAL_PROC_STAT
paraski33% cat /proc/cpuinfo
  processor       : 0
  vendor_id       : GenuineIntel
  cpu family      : 15
  ...
paraski33% cat /proc/meminfo
  MemTotal:       498944 kB
  MemFree:        457936 kB
  ...
```

Capacités : autres exemples

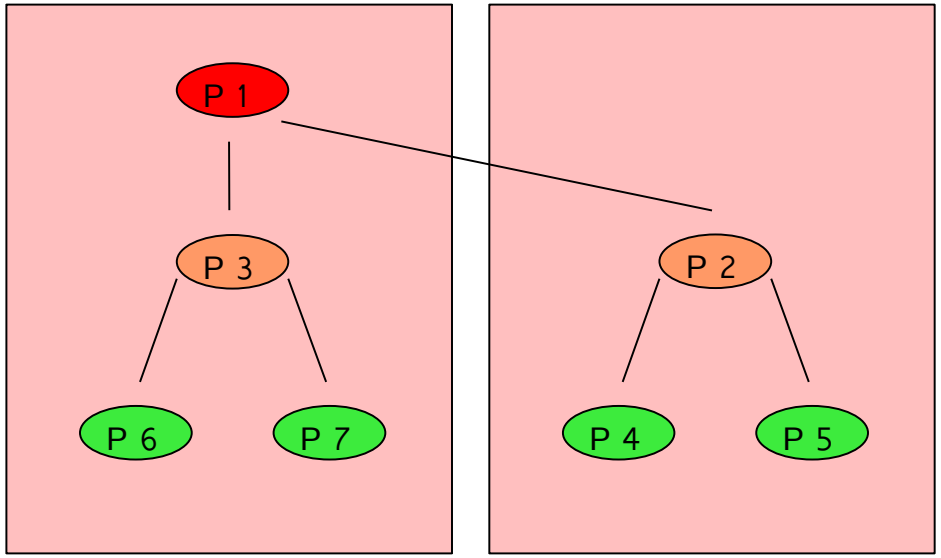
```
void main(void)
{
	int nb_sons = 0;
	int depth = 0;

	while (nb_sons != 2 && depth < 2) {
		if (fork())
			nb_sons++;
		else {
			depth++;
			nb_sons = 0;
		}
    }
}
```



Capacités : exemple 2

```
void main(void)
{
	int nb_sons = 0;
	int depth = 0;

	while (nb_sons != 2 && depth < 2) {
		if (fork())
			nb_sons++;
		else {
			depth++;
			nb_sons = 0;
		}
	}
}
```

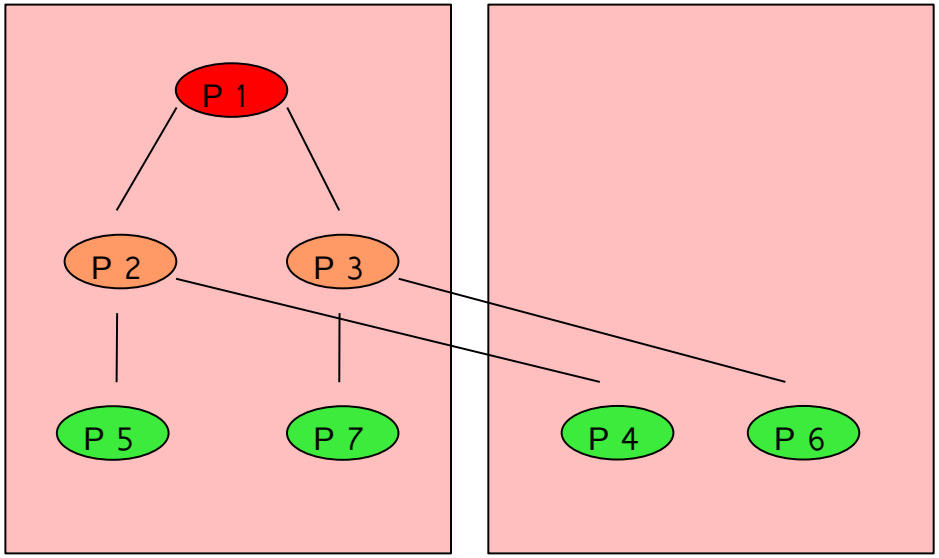
```
paraski33% krg_capset -e +DISTANT_FORK
paraski33% ./fork-test
```



Capacités : exemple 3

```
void main(void)
{
	int nb_sons = 0;
	int depth = 0;

	while (nb_sons != 2 && depth < 2) {
		if (fork())
			nb_sons++;
		else {
			depth++;
			nb_sons = 0;
		}
    }
}
```

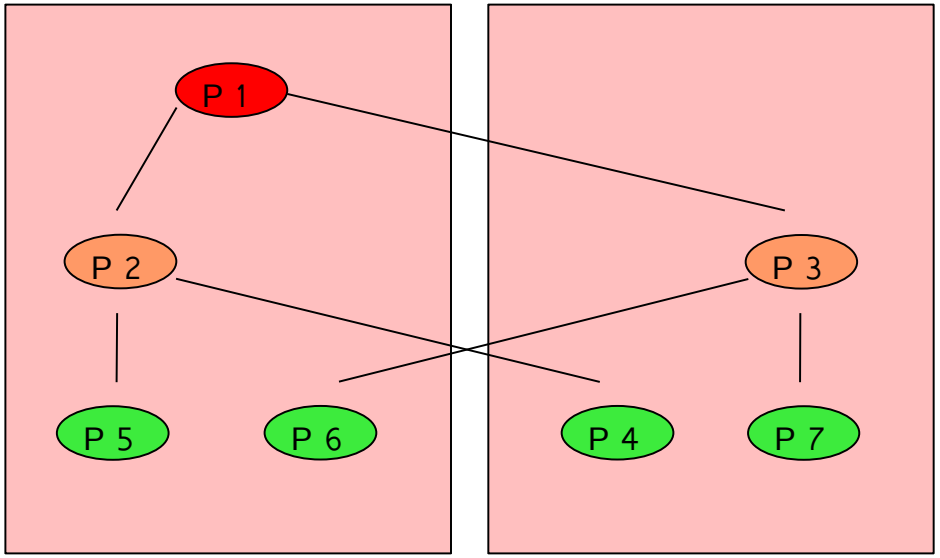
```
paraski33% krg_capset -d +DISTANT_FORK
paraski33% ./fork-test
```



Capacités: exemple 4

```
void main(void)
{
	int nb_sons = 0;
	int depth = 0;

	while (nb_sons != 2 && depth < 2) {
		if (fork())
			nb_sons++;
		else {
			depth++;
			nb_sons = 0;
		}
	}
}
```

```
paraski33% krg_capset -e +DISTANT_FORK
paraski33% krg_capset -d +DISTANT_FORK
paraski33% ./fork-test
```



Plan

- Généralités
- Gestion globale des processus
- Gestion globale de mémoire
- Gestion globale des fichiers
- Gestion globale des communications
- Paramétrage du système
- Administration de la grappe
- Gestion des défaillances
- Expérimentations
- Conclusion

Administration de la grappe

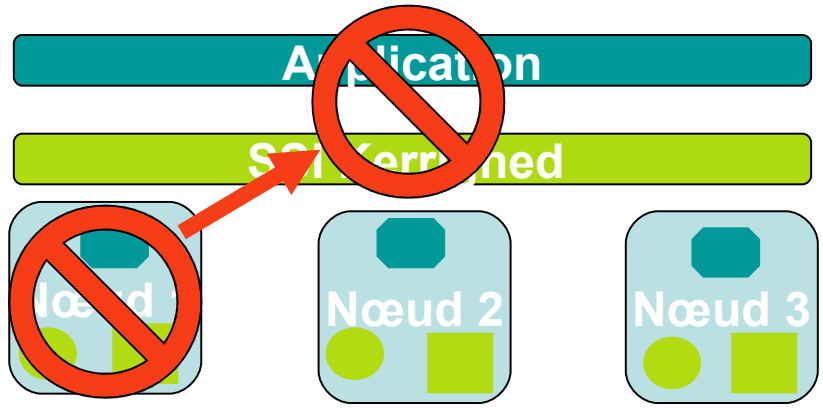
- Démarrage de la grappe
  - `krgadm cluster start`
- Arrêt de la grappe
  - `krgadm cluster stop`
- Ajout de nœuds à chaud
  - `krgadm nodes add -n16:18`
- Arrêt de nœuds à chaud
  - `krgadm nodes del -n21:24`

Plan

- Généralités
- Gestion globale des processus
- Gestion globale de mémoire
- Gestion globale des fichiers
- Gestion globale des communications
- Paramétrage du système
- Administration de la grappe
- Gestion des défaillances
- Expérimentations
- Conclusion

Impact d'une défaillance

- Défaillance d'applications en cours d'exécution
  - Perte des processus exécutés sur le nœud défaillant
  - Perte de données stockées en mémoire ou sur disque sur le nœud défaillant
- Perte d'information et de serveurs système
  - Dans un SSI, les nœuds coopèrent pour doner l'illusion d'une seule instance du systèe Linux



Points de reprise d'applications

- Sauvegarde et restauration de l'état des applications
  - Famille de processus répartis sur les nœuds d'une grappe
    - Père et fils (fork)
  - Processus multithreadés
  - Processus communiquant en utilisant des IPC
    - Segments de mémoire system V (mémoire partagée)
    - Files de messages
    - Sémaphores
- Optimisations
  - Sauvegarde incrémentale de processus
    - Seule les données modifiées depuis le dernier point de reprise sont sauvegardées
  - Sauvegarde efficace des fichiers modifiés par l'application (support dans le système de fichier distribué kDFS)
- Travaux en cours ou futurs
  - Points de reprise d'applications communiquant par messages
    - Sockets inet, sockets unix, pipes, FIFO, char devices
  - Prise en compte des fichiers de l'application dans la sauvegarde et restauration de points de reprise

Haute disponibilité du système Kerrighed

- Travaux de recherche en cours
  - Fiabilisation du système de communication utilisés par les services distribués de Kerrighed
  - Service générique de gestion des défaillances
    - Détection des défaillances
      - Vision globale pour tous les nœuds de la grappe
    - Notification des différents services distribués sur tous les nœuds
  - Chaque service doit prévoir les actions à exécuter lorsqu'une défaillance survient
    - Nettoyage de structures de données pour prendre en compte les informations et entités perdues suite à la défaillance
    - Arrêt d'applications dans certains cas

Plan

- Généralités
- Gestion globale des processus
- Gestion globale de mémoire
- Gestion globale des fichiers
- Gestion globale des communications
- Paramétrage du système
- Administration de la grappe
- Gestion des défaillances
- Expérimentations
- Conclusion

Expérimentations

- Tests de non régression exécutés quotidiennement
  - KLTP (version étendue de la suite de tests standard utilisées par valider l'interface Posix sur Linux)
- Tests de passage à l'échelle et de support des configurations matérielles
  - 256 nœuds
  - Nœuds SMP
  - Processeurs multi-cœurs
  - Processeurs 32 et 64 bits
- Tests avec des applications réelles fournies par des partenaires indistriels ou académiques (EDF R&D, DGA, ONERA-CERT, Cap Gemini, Xlab, EADS, SAP …)
  - Applications séquentielles et parallèles
  - OpenMP
  - MPI

Plan

- Généralités
- Gestion globale des processus
- Gestion globale de mémoire
- Gestion globale des fichiers
- Gestion globale des communications
- Paramétrage du système
- Administration de la grappe
- Gestion des défaillances
- Expérimentations
- Conclusion

Conclusion

- Version courante Kerrighed 2.3.0 fondée sur le noyau Linux 2.6.20
  - Licence GPL-2
  - Packagée pour Mandriva et RedFlag Linux
  - http://www.kerrighed.org
  - Nouvelle version en novembre 2008 (SC'08)
  - Portage vers de nouvelles versions du noyau à venir (noyau le plus récent)
- Offre commerciale Kerlabs
  - Support garanti
  - Paramétrage du système
  - Développements spécifiques
- Travaux de recherche en cours ou futurs
  - Amélioration du système de fichiers distribués et parallèle kDFS
  - Amélioration des mécanismes de tolérance aux fautes pour les applications
  - Haute disponibilité du système
  - Politiques d'ordonnancement global
    - Gestion de l'énergie
    - Prise en compte des entrées/sorties disque
- Contributeurs (utilisateurs ou développeurs) bienvenus dans la communauté

Informations et contacts

- Information
  - http://www.kerrighed.org
  - http://www.kerlabs.com
  - http://www.xtreemos.eu (LinuxSSI)
  - http://www.irisa.fr/paris (publications)

- Contacts
  - Christine Morin, INRIA Rennes Bretagne Atlantique
    - Christine.Morin@inria.fr
  - Dominique Loucougain et Pascal Gallard, Kerlabs
    - Dominique.Loucougain@kerlabs.com
    - Pascal.Gallard@kerlabs.com

KERLABS