Deploiement d'un cluster Altix ICE

ANGD 7 Octobre 2009

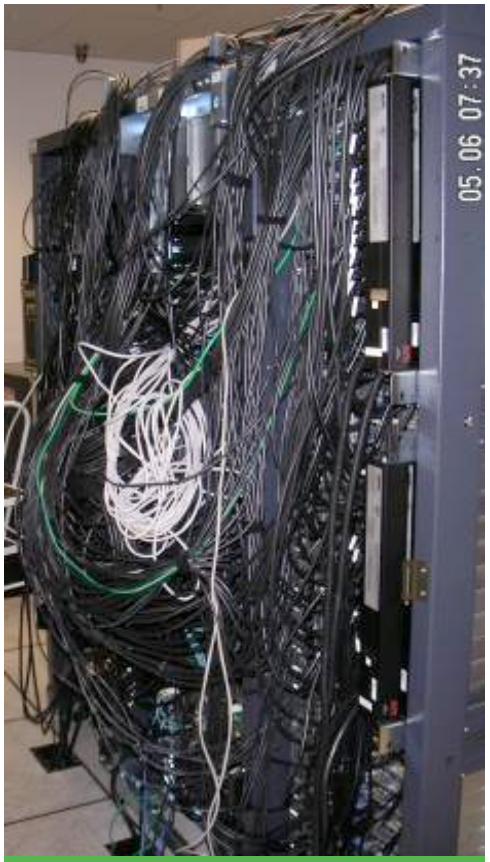
Sébastien Piechurski

Présentation de l’environnement Altix ICE

Présentation de l'ICE 8200

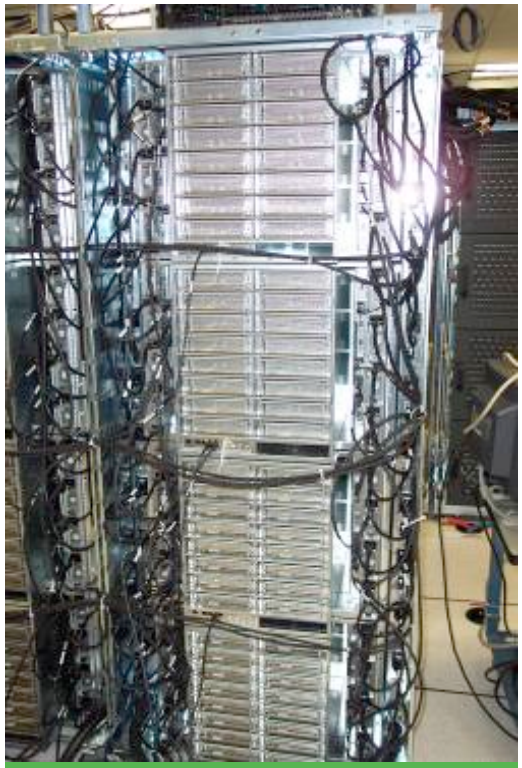
- ICE = Integrated Cluster Environment
- Racks haute densité
  - Jusqu'à 64 lames / rack => 512 cores
- Basé sur les chipset Intel®
  - Dual et quad core
- Jusqu'à 96Go de mémoire par lame
- Réseau Infiniband 4x DDR bi-plan intégré
- Alimentation et refroidissement intégré
  - Portes refroidissantes en option (refroidissement par eau)
  - Double économie d'énergie
- Architecture hiérarchique et diskless

Présentation de l'ICE 8200



Cluster traditionnel

- Réseau intégré
  - IRU = 128 cores sans cables
  - Double réseau Infiniband + 1 réseau d'administration (Ethernet)
- Alimentations et refroidissement redondant et remplaçable à chaud



Altix ICE

Détail d'un rack ICE

- Chaque rack 42U (30" L x 40" P) comprend:
  - 4 IRUs avec 16 lames bi-socket
  - 128 sockets Intel Xeon DP (jusqu'à 512 cores)
  - 48 ports Infiniband DDR 4x pour connexions externes
  - 1 contrôleur de rack – services de démarrage
- En option pour les grandes configurations: des unités de refroidissement liquide (portes froides)

sgi

Architecture hiérarchique

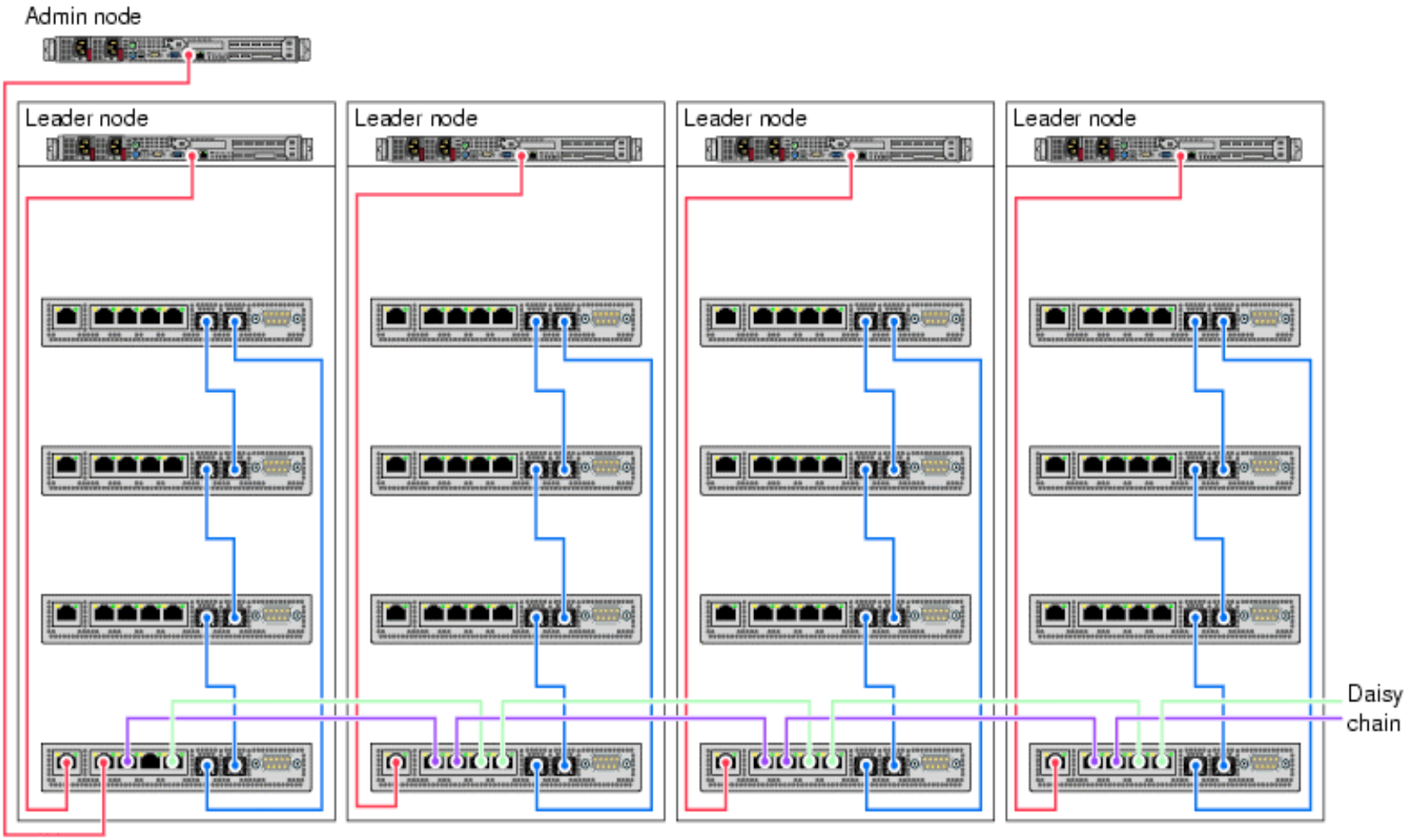
- Matériel de gestion hiérarchique:
  - Systèmes de gestion
    - Nœud d'administration système (1 par cluster)
      - Point central pour le(s) adminitrateur(s)
    - Rack leader (1 par rack)
    - CMC (Chassis Management Controller) (1 par IRU)
    - BMC (Baseboard Management Controller) (1 par lame)
  - Nœuds de service
    - Fournit les services utilisateurs (nœuds de login)
      - Point central pour les utilisateurs du cluster
    - Autres services
      - Nœud de stockage
      - Nœud gestionnaire de travaux
      - Nœud passerelle

sgi

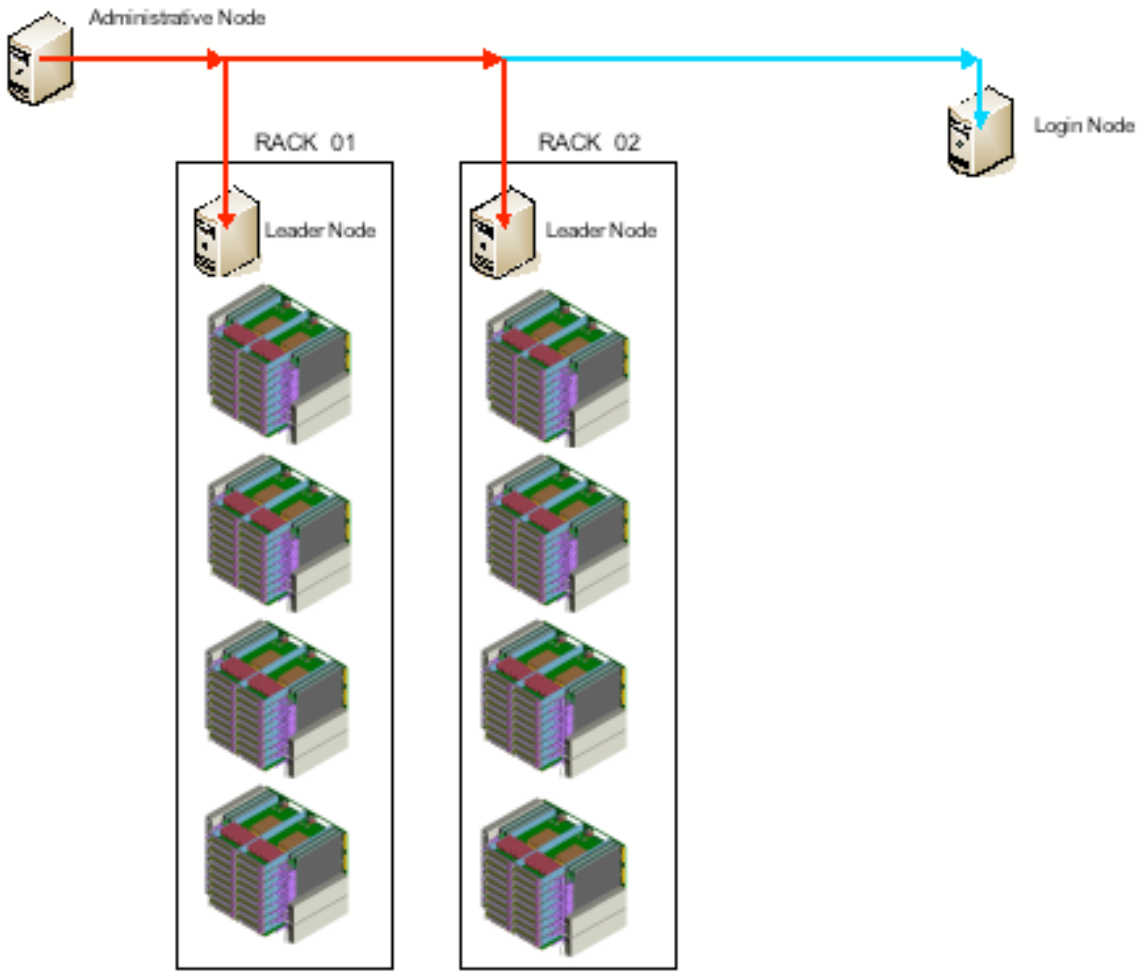
Les réseaux sur l'altix ICE

- Infiniband
  - ib0 utilisé pour la communication interne aux jobs (MPI, autres messages, …)
  - ib1 utilisé pour le traffic lié au stockage
  - MPT (MPI SGI) sait utiliser les 2 simultanéments pour augmenter les débits et réduire les latences
  - Aucun n'est utilisé pour l'administration
- Ethernet Gigabit
  - N'est pas utilisé directement par les utilisateurs et les applications
  - Vue matérielle
    - Prévu pour réduire le nombre de câbles et faciliter les extensions
  - Vue logicielle
    - Géré par des VLans
    - Prévu pour la scalabilité
    - Non global

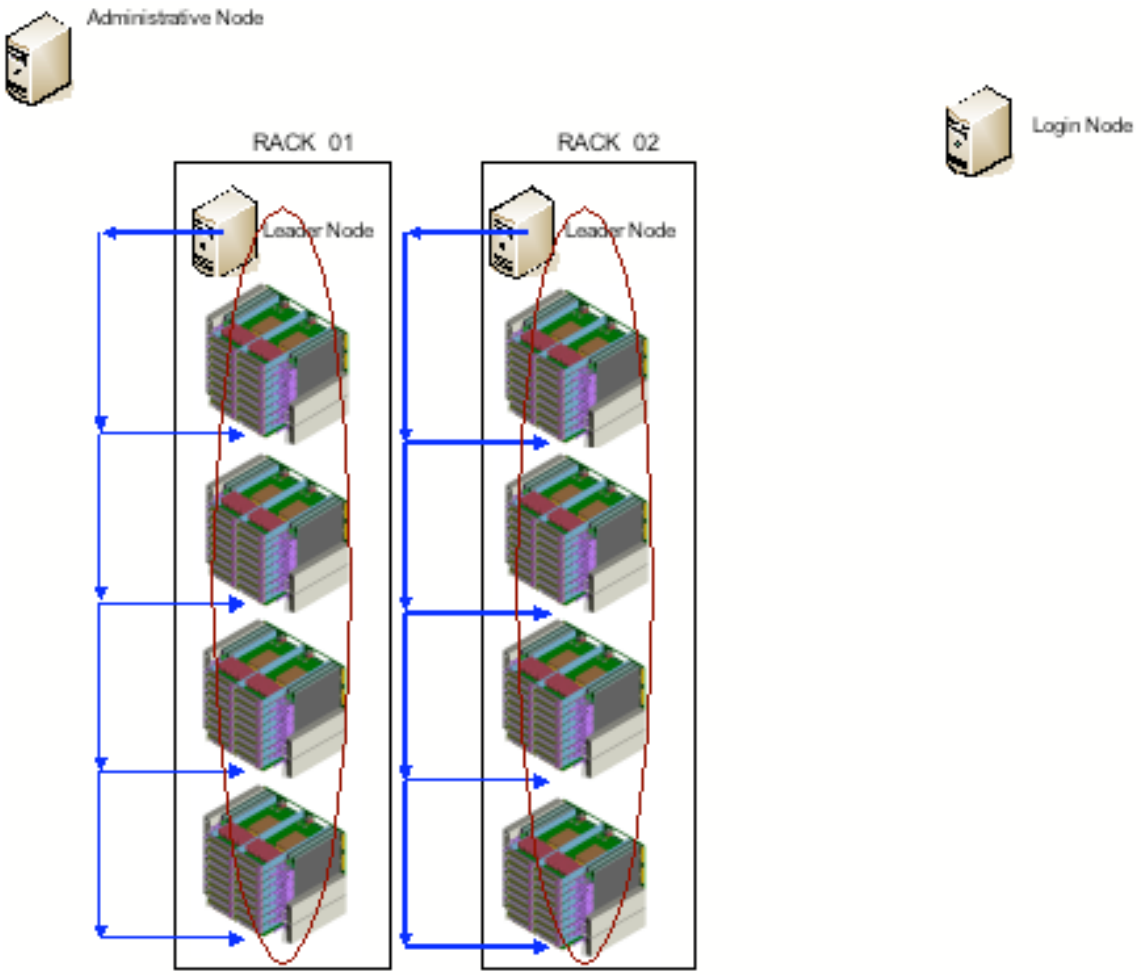
Réseau Gigabit: le câblage



Le réseau Gigabit: VLAN_HEAD



Le réseau Gigabit: VLAN_GBE et VLAN_BMC



Nœud d'administration système

- Livré installé avec la dernière version de Tempo
  - En cas de réinstallation, utiliser le DVD « Altix ICE Admin Controller Install »
- Découvre et provisionne les autres nœuds:
  - Crée les images pour les lames
  - Découvre et installe les racks
  - Découvre et installe les nœuds de service
- Contrôle le demarrage/arrêt matériel des autres nœuds
- Point central pour la surveillance (Ganglia, ESP, PCP)
- Serveur DNS primaire pour l'ensemble du cluster
- Détient la base de données de Tempo

sgi

Rack leader

- 1 par rack
- N’est pas directement accédé par les utilisateurs et administrateurs (non requis)
- Découvre les CMCs (et IRUs), les lames et leur BMC
- Est le serveur de boot pour les lames du rack
- Point central de surveillance pour le rack
  - Collecte et stocke les données des lames, CMCs et BMCs
  - ESP et PCP retransmet des infos au nœud d’admin
- Sert de cache DNS pour les lames du rack
- Tourne le « subnet manager » pour les réseaux IB

Les CMCs et BMCs

- CMC = Chassis Management Controller
  - 1 par IRU
  - Switch ethernet intégré
  - Configure les VLans
  - Identifie à quel IRU il appartient
  - Découvre les lames de l'IRU et fournit la liste sur demande
  - Contrôle l'alimentation des lames
  - Collecte les infos sur les ventilateurs, l'énergie et la température
- BMC = Baseboard Management Controller
  - 1 par lame
  - BMC standard, interface IPMI
  - Fournit les fonctionnalités de console (Serial Over Lan), contrôle d'alimentation, surveillance, …

sgi

Nœuds de service

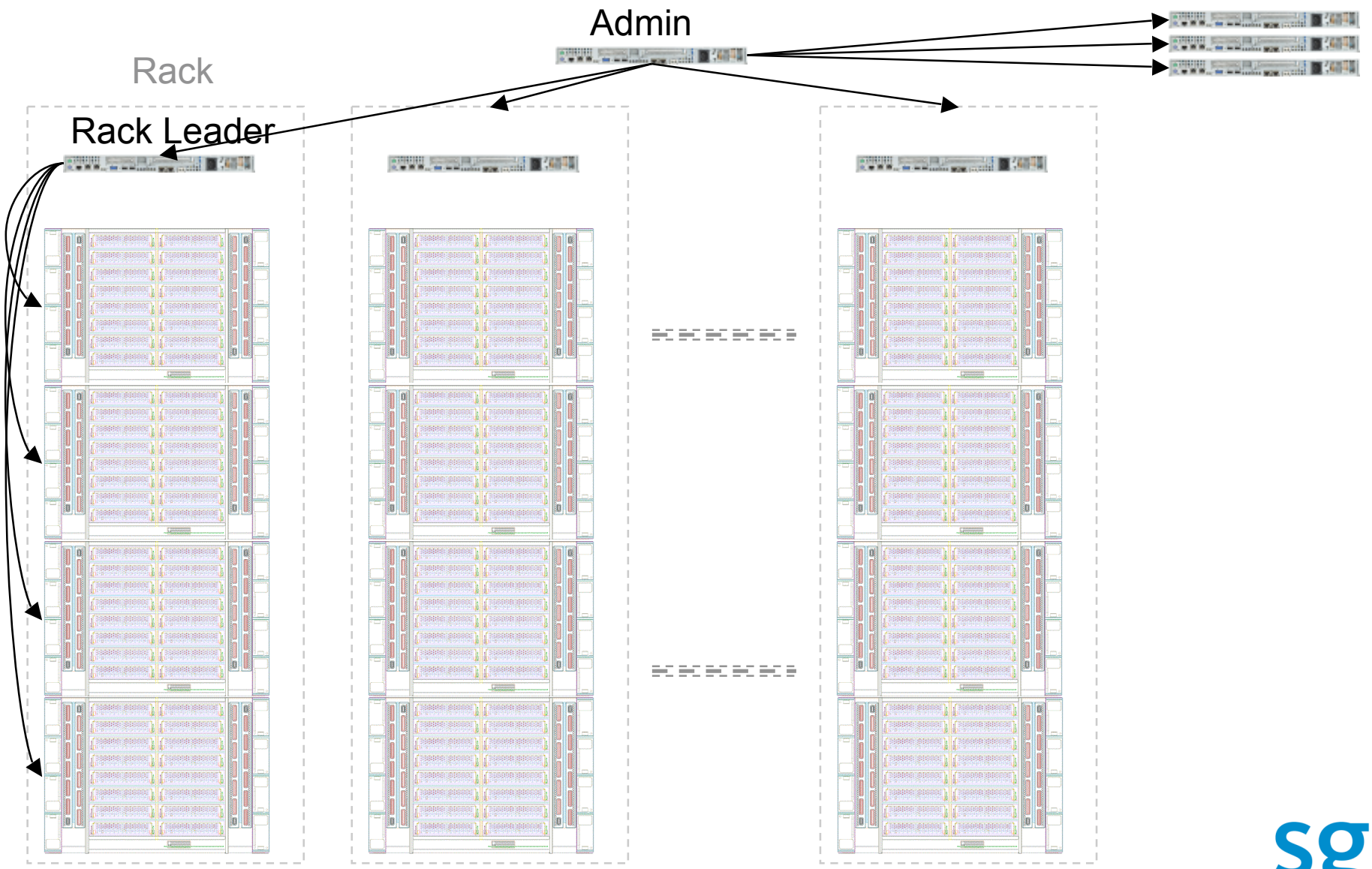
- Nœud 1U ou famille SGI XE
- Nœuds de login
  - Point central pour les utilisateurs
  - Souvent utilisé pour le développement, la compilation
  - Sert également à soumettre les jobs
- Serveur de batch
  - Altair PBS, Torque, SGE, etc …
- Nœud passerelle
  - NAT, NIS, NFS, LDAP, etc …

sgi

Nœuds de stockage

- Serveur NFS
- Serveur Lustre
- Destiné à fournir un volume de stockage commun pour tous les nœuds du cluster (home, scratch, …)

Evolutivité



SGI Tempo

Tempo

- Logiciel central à la gestion du cluster
  - Basé sur des programmes open-source
    - Oscar, Systemimager, Ganglia, Conserver, pdsh, MySQL
  - Complété par des programmes SGI
    - PCP, ESP
    - cpower, discover, blademon
- Permet:
  - Gestion des images systèmes (création, mise à jour, déploiement, ...)
  - Reconfiguration automatique des systèmes et composants logiciels lors d'ajout et suppression de nouvelles machines (DHCP, DNS, etc …)

sgi

Intégration matérielle

- Tempo a été conçu spécialement pour l'environnement ICE
- Sait gérer les machines suivantes:
  - ice-csn (machines 1U)
  - Famille XE: 210, 240, 250, 270, 310, 320, 350, etc..
- Ainsi que les switches Infiniband:
  - Internes
  - Externes

sgi

Installation rapide et facile

- Installation du nœud d'administration (OS + Tempo combinés)
- Assistant d'installation (configure-cluster)
- Découverte quasi-automatique de multiples racks
  - Allumage manuel des rack leaders et nœuds de service (un par un) pour découvrir les adresses MAC
  - Redécouverte automatique en utilisant une liste d'adresses MAC
- Découverte automatique des CMCs et des lames
- Gestion automatisé du remplacement à chaud des lames

sgi

Administration système efficace

- Contrôle des systèmes et de l'alimentation hiérarchique
- Pas d'influence sur les nœuds de calcul
  - Surveillance des nœuds très légère
  - Collection de données optimisée via le push PCP
- Scalabilité démontrée jusqu'à 6400 lames / 51200 cores (Nasa)
- Surveillance scalable
  - Cluster PCP (plus de 40 métriques par nœud par défaut, 1000 disponibles)
  - Logs des données BMC SEL (System Event Log)
  - Logs syslog
  - Logs des console

sgi

Gestion de système unique

- Multiboot en cascade
- Diskless
  - Partition root via NFS
    - En grande partie en lecture seule
    - /root, /etc et /var en lecture/écriture
  - Partition root via tmpfs (entièrement en mémoire)
- Gestion des mots de passe BMC

sgi

Provisioning simple et efficace

- Gestion des images
  - Support de dépôts multiples
  - Gestion de différents types d'image et de versions
  - Duplication d'image
  - Personnalisation par le client
  - Supporte les mises à jours (sans perte des personnalisations propres au site)
- Provisioning dynamique – sélection d'une image par nœud
- Distribution des images
  - Hiérarchique
  - Sans disque pour les lames
  - Avec disque pour les rack leaders et nœuds de service
  - Mises à jour incrémentales à chaud
- Supporte les distributions SLES 10/11 et RHEL 5

sgi

Maintenance facilitée

- Possibilité d'échange des racks leaders / nœud d'admin à froid
- Surveillance
  - Ganglia
  - ESP
  - PCP
- Outils:
  - Console disponible pour tous les nœuds
  - Outil de vérification des interconnects (cablâge)
  - Interface unique de regroupement de configuration « tempo-info-gather »

sgi

Plus de détails

Vue globale de l'installation

- On commence avec le nœud d'administration
  - Installation depuis le DVDTempo
  - Personnalisation basique du système (réseau, …)
  - Personnalisation de l'ICE via configure-cluster
  - Personnalisation des images des lames et nœuds de service (montages, fichiers de config…)
- Installation des rack leaders
  - On lance la commande « discover », puis on allume les racks leader un par un
  - Découverte automatique et installation des rack leaders
  - Découverte automatique des CMCs
  - Découverte automatique et provisioning des lames
- Installation des nœuds de service
  - On lance la commande « discover », puis on allume les nœuds de service un par un
  - Découverte automatique et installation des nœuds de service

Stratégie de démarrage diskless

- Le rack leader est le serveur de démarrage
  - Un maximum de 64 lames peuvent démarrer depuis le rack leader
- 2 possibilités pour le volume root
  - NFS
  - TMPFS
- /tmp est un petit tmpfs
  - Taille configurable
- Swap minimum
  - Via iscsi depuis le rack leader
  - Facilement désactivable
  - Si réellement requis, conseillé de le déporter vers un nœud de stockage
- Pour les autres systèmes de fichiers comme applications, home, données utilisateurs
  - Utilisation fortement recommandée d'un nœud de stockage (NFS, Lustre, Panasas, …)

sgi

Stratégie NFS

- Utilise la capacité de Linux à utiliser un volume root NFS
- Minimise l'utilisation mémoire sur les nœuds de calcul sans disque
- Le volume root est monté en lecture seule et partagé par toutes les lames du rack depuis le rack leader
  - Rentre dans le cache du rack leader
- Chaque lame a sa propre zone privée en lecture / écriture
  - Regroupe /var, /etc et /root
  - Créé par cimage au moment du transfert d'image vers le rack leader
  - Taille limitée par quota XFS pour éviter de remplir le disque du leader

sgi

Stratégie TMPFS

- Le volume root entier réside en mémoire dans un montage tmpfs
- Utilise la même infrastructure sur le leader que le NFS
  - Possible facilement de passer d'une stratégie à l'autre
  - Fonctionnement:
    - Un tarball du root est envoyé via multicast
    - Les différences par lames sont synchronisées individuellement via rsync
- Les modifications ne sont pas gardées entre 2 redémarrages
- Plus de mémoire requise:
  - L'image doit être inférieure à la moitié de la mémoire disponible sur la lame
  - L'image par défaut fait 1,1Go
  - Mais l'image est personnalisable

sgi

Comparaison des 2 stratégies

Root NFS

- Modifications persistentes
- Requiert moins de mémoire
- Assez rapide
- Dépendant de l'état du leader
- Les modifications systèmes doivent être faites sur le leader

Root TMPFS

- Modification non persistentes
- Requiert plus de mémoire
- Très rapide
- Les lames peuvent tourner sans le leader
- Des programmes peuvent être installés directement sur les noeuds

sgi

Image système des lames

- SGI fournit une image par défaut basé sur la distribution standard
  - Liste de RPMs choisie par SGI
  - Cette image suffit pour la plupart des cas
- Plusieurs images sont supportées
  - Cloner ou créer de nouvelles de zéro
  - Personnaliser pour le site
  - Peut-être personnaliser pour un utilisateur ou une application
- On passe par l'interface d'administration pour définir quelle image démarrer la prochaine fois

sgi

Contrôle d'alimentation - cpower

- S'applique au cluster complet ou un composant en particulier
  - --system
  - --rack, --iru, --node
- Inclus les opération matérielles et logicielles
  - --on, --off, --reset, --cycle
  - --boot, --reboot, --shutdown, --halt
  - --status
- cpower tient compte de la hiérarchie
  - Allume les leaders avant les IRUs, les IRUS avant les lames, etc…
- Une seule commande pour allumer le cluster complet
  - cpower --system --on

sgi

Conserver

- Gère la console de tous les nœuds (sauf admin)
- La configuration gère la hiérarchie
  - Le démon tourne sur les leaders pour gérer la console des lames
  - Le démon tourne sur l'admin pour gérer les leaders et les nœuds de service
  - Le démon maintient la connexion avec « ipmitool »
- Toutes les consoles sont accessibles depuis l'admin
- Une connexion en écriture + multiples connexions en lecture seule possible
- Possibilité de « voler » la connexion en écriture
- Toutes les sorties consoles sont loggées
- Plus d'infos sur http://www.conserver.com

sgi

Fiabilité

- Design du système
  - Matériel SGI de qualité
  - Alimentations redondantes partout
  - Les nœuds de services et les rack leaders ont des disques en mirroir (RAID 1 matériel)
  - Les lames sont sans disque et insérables à chaud
- Une attention particulière sur la détection de panne et sa réparation:
  - ESP, PCP, Ganglia
  - Découverte et intégration automatique des lames dans les programmes sus-cités
  - Remplacement à froid pour l'admin et les rack leaders
    - Les disques sont mirrorés
    - Si autre chose tombe en panne, il suffit de remplacer la machine et bouger les disques

Outils de diagnostique

- Surveillance
  - ESP
  - PCP
  - Ganglia
- Outils linux standards
  - Syslog
  - Consoles
  - Kdump
  - SysRq
- Diagnostique matériel
  - Memlog
  - Diagnostique Infiniband
  - Diagnostique CPU/mémoire
- Spécifique à Tempo:
  - Outil de vérification d'interconnect (ivt)
  - Tempo-info-gather

sgi

Support des systèmes

- SLES 10
- SLES 11
- RHEL 5
- Autres options (pas de support officiel)
  - CentOS 5 et Scientific Linux 5
  - Certaines version de Fedora
  - Certaines version d'OpenSUSE

L’installation en pratique

Installation d'une ICE de A à Z

- Installation du nœud d'administration (~40 min)
- Configuration du cluster (~40 min)
- Création d'une image personnalisée (~15 min)
- Déploiement des rack leaders (~15 min)
- Démarrage des lames (~3 min)
- Configuration du réseau Infiniband (~3 min)
- Déploiement des nœuds de service (~15 min)

Total temps de deploiement: ~2 heures 15 min

sgi

Installation du nœud d'administration

- Insérer le DVD de Tempo dans le lecteur
- Allumer le nœud d'admin
- Choisir le nombre de slots avec les options
  - re_partition_with_slots
  - install_slot
- Valider
- Pause (environ 45 minutes)

Configuration du nœud d'administration

- Paramètrage système standard:
  - Réseau
    - Adresse IP externe
    - Masque de sous réseau
    - Passerelle
    - Serveurs de nom
  - Date et heure
  - Mot de passe root
  - Modèle d'authentification

Configuration du cluster

- Importation des packages depuis les CD / ISOs
- Création des dépôts standard (SLES, ProPack)
- Définition des sous réseaux internes
  - Le défaut doit convenir à la plupart des installations
- Définition du nom de domaine interne
- Configuration NTP
- Génération des images standards
- Configuration des serveurs de noms

Création des images système personnalisées

- Dépôt(s) personnalisé(s)
  - Possibilité d'ajouter ses propres RPMs
- Personnalisation d'une liste de paquets
  - Peut inclure les paquets propres au site
- Scripts post-install personnalisés
  - Pour apporter des modifications aux images
  - Conseillé pour garder la traçabilité des modifications
  - Réutilisable en cas de mise à jour

sgi

Découverte des rack leaders

- Lancement de « discover »
- Allumage des rack leaders
- Installation de l'image des rack leaders
- Installation de l'image par défaut des nœuds de calcul
- Découverte automatique, intégration et allumage des nœuds de calcul

Configuration des réseaux Infiniband

- Via configure-cluster
- Choix de la topologie du réseau
  - Dépendant du câblage physique
- Choix des machines faisant tourner le subnet manager
  - N'importe quel rack leader
  - Fonctionne par paire pour la redondance (si possible)
  - Une paire par rail (aussi si possible)

sgi

Gestion quotidienne

- Logs disponible dans un point unique:
  - /net/*/var/log/messages
- Logs de toutes les consoles disponibles
  - /net/*/var/log/consoles/*


sgi
accelerating
results™