Module 8: Infrastructure pour le calcul de haute performance

INF8601: Systèmes informatiques parallèles

Michel Dagenais

Calcul de haute performance

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Ordinateurs gigantesques occupant des dizaines, centaines ou plus de chassis.
- Infrastructure physique: édifice pour la sécurité et contre les intempéries, alimentation électrique, climatisation.
- Chassis, boîtiers et composants électroniques.
- Câblage de réseautique.
- Surveillance et commande de système et du réseau.
- Entretien logiciel et matériel.
- Support aux utilisateurs.

Infrastructure physique

- Edifice à l'épreuve des tremblements de terre, émeutes et inondations (anciennes centrales téléphoniques).
- Centre mobile dans un conteneur.
- Bâtiment rudimentaire avec ventilation extérieure pour minimiser le coût.
- Alimentation électrique de grande puissance avec plus d'un circuit d'entrée.
- Eau ou air de refroidissement. (Pays nordique?).

Climatisation

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Chaque watt consommé pour le calcul compte double, il génère de la chaleur et doit être évacué par climatisation.
- Air forcé à travers le chassis (de bas en haut, d'avant à l'arrière, panneaux refroidissants à eau).
- Compromis sur la température entre l'usure du matériel électronique et le coût de climatisation.
- Utiliser l'air extérieur, accepter une température plus haute et économiser beaucoup.
- Environnement sans humain, non éclairé...

Alimentation électrique

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Alimentation du secteur pour chaque ordinateur, avec bloc d'alimentation redondant dans chaque boîtier. Alimentation de secours avec piles et génératrice.
- Bloc d'alimentation partagé pour un chassis ou plus, avec ou sans redondance. Alimentation de secours en courant continu.
- Ecolotique (Green Computing). Performance en FLOP/Watt, Blue Gene, serveurs ARM.

Noeuds individuels

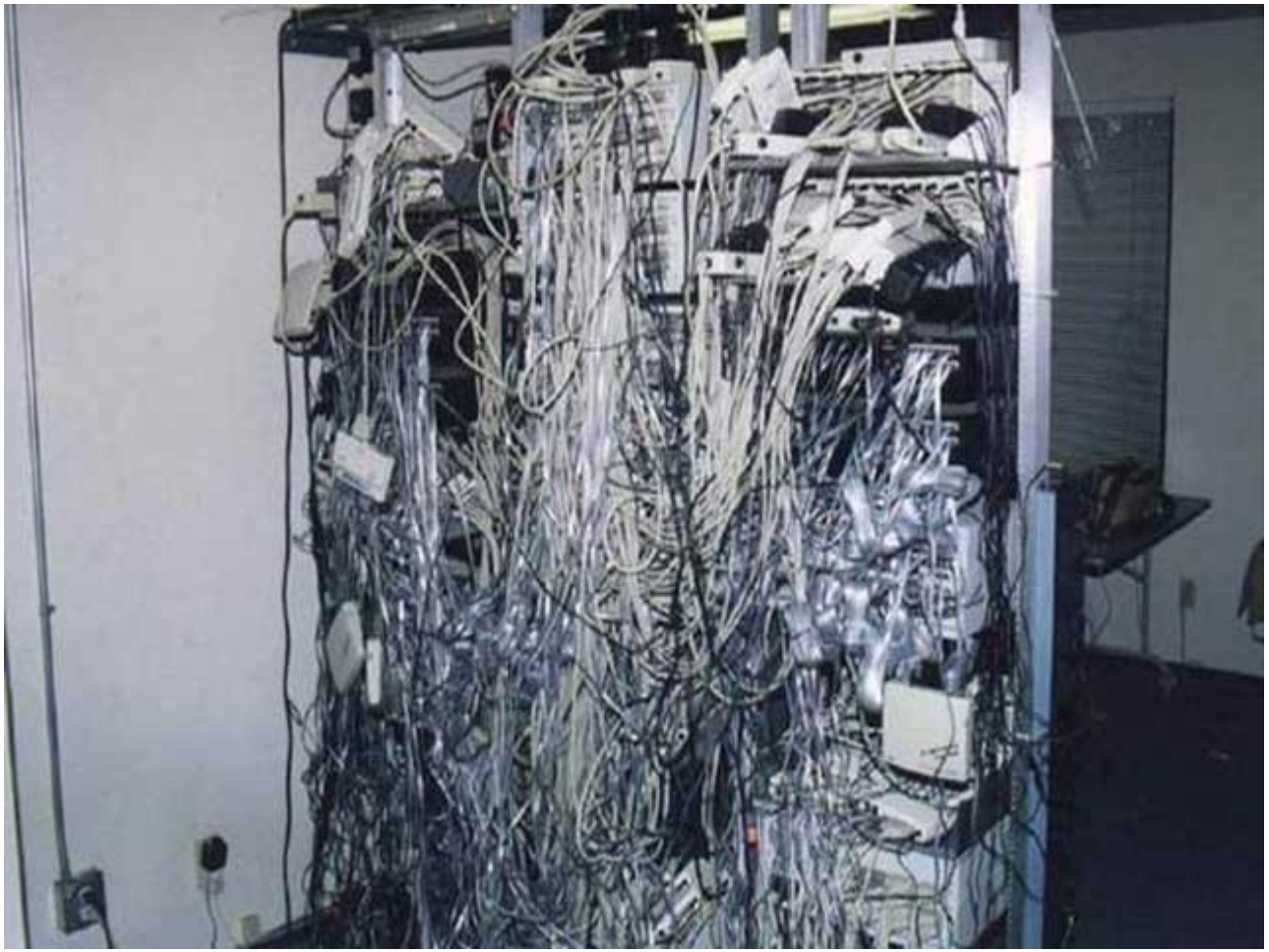
- Boîtiers 1U ou 3U pour un noeud, ou grands boîtiers avec un noeud par carte verticale (blade). Matériel haut de gamme avec chaque composant très rapide ou très fiable.
- Organisation sur mesure ou utilisant les composantes de grand volume pour réduire les coûts (boîtiers en tôle pliée compacts, cartes électroniques vendues à grand volume, disques SATA...). Facebook Open Hardware, Google.

Réseautique

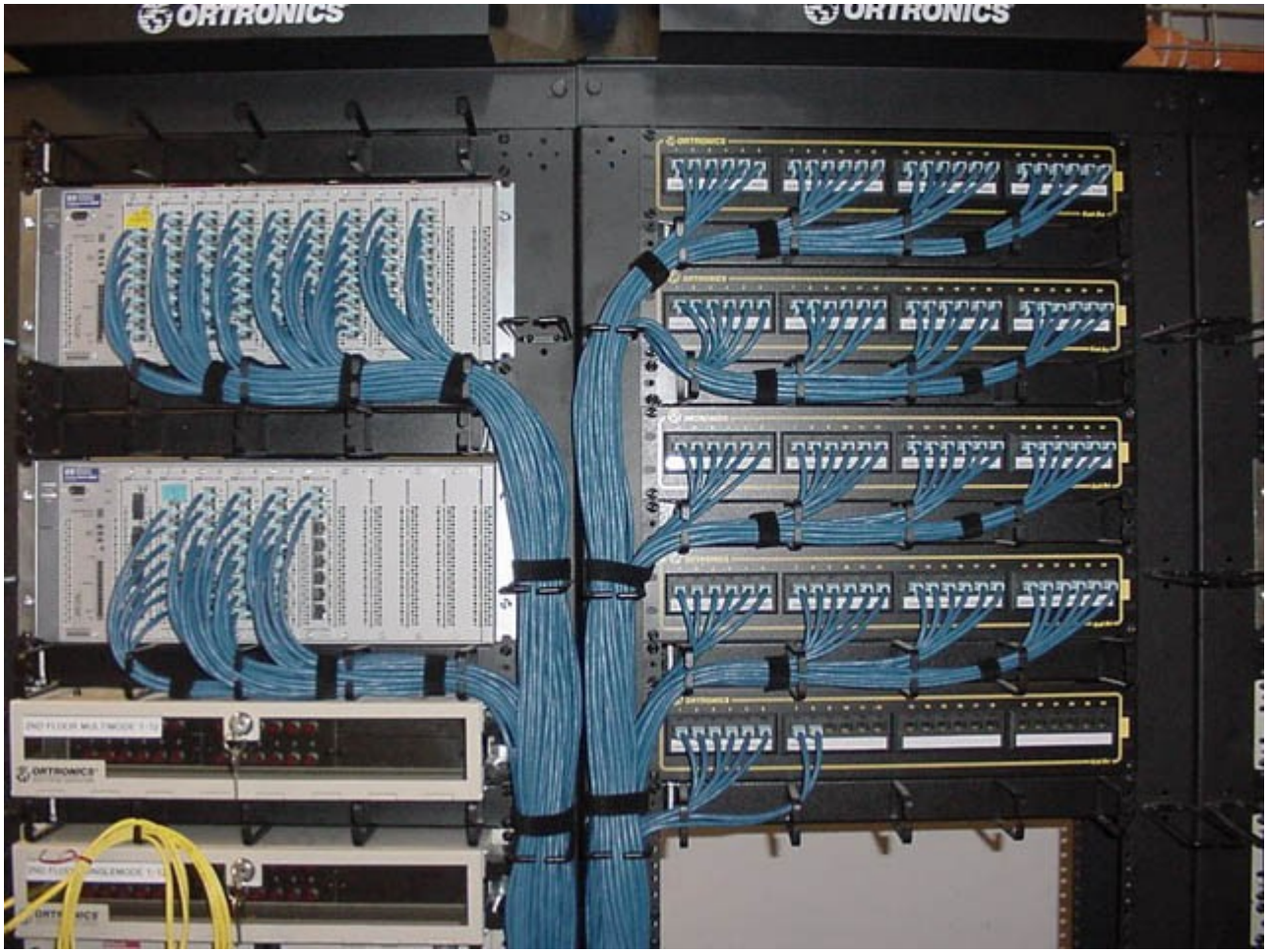
ÉCOLE POLYTECHNIQUE MONTRÉAL

- https://en.wikipedia.org/wiki/List_of_interface_bit_rates
- Réseaux spécialisés à faible latence comme InfiniBand, 2 à 600 Gbps, (qui supplante Myrinet, 640Mbps à 10Gbps).
- Réseaux performants moins chers (Ethernet 10Mbps à 400Gbps).
- Connexions point à point pour faible latence versus commutées.
- Graphe pleinement versus partiellement connecté.

Le câblage



Le câblage



Surveillance de système et réseau

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Agent SNMP ou l'équivalent sur chaque routeur et noeud, qui permet de lire différentes métriques (température du processeur et du boîtier, taux d'utilisation des ressources CPU, disque ou réseau) ou d'envoyer des alertes.
- Console de gestion qui teste les services régulièrement (ping, accès fichier, accès Web...).
- Journal centralisé des erreurs sur chaque noeud (syslog).
- Nagios, OpenNMS, Ganglia, IBM Tivoli, HP Network Management Center.

Commande de système et réseau

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Commande des routeurs par SNMP.
- Commande de l'alimentation et remise à zéro à distance pour les noeuds par KVM-IP ou BIOS UEFI, redirection du clavier/souris et de l'affichage de la console.

Entretien matériel

- Remplacer les disques, blocs d'alimentation ou cartes électroniques.
- Problèmes subtils causés par du matériel en apparence identique mais pas complètement compatible, interférence, chaleur...
- Contrats de service 4 heures ou 24 heures, pièces de rechange garanties, ou entretien par soi-même.

Entretien logiciel

- Mises à jour de sécurité et correction de bogues.
- Mises à jour de versions.
- Ajout de logiciels.
- Vérification de l'intégrité des systèmes, détection d'intrusion, détection d'anomalies.

Support aux usagers

- Documentation de l'environnement.
- Séances de formation.
- Aide spécifique aux utilisateurs.
- Réponse aux questions.
- Résolution de problèmes.

Les grappes de calcul

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Budget important aux 3 à 5 ans.
- Infrastructure physique importante.
- Assurer une pleine utilisation.
- Allocation de ressources entre les usagers.
- Accès à distance, configuration flexible (choix de librairies, outils, version du noyau de système d'exploitation).
- Demande de maintenir une équipe importante de spécialistes informatiques qui coûtent cher.
- Virtualisation? Infonuagique? Confidentialité?

Infrastructure logicielle d'une grappe

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Un ou quelques noeuds de gestion, un grand nombre de noeuds parallèles.
- Procédure d'installation des noeuds (Linux) par réseau et d'ajout ou retrait de noeud.
- Exécuter des commandes sur l'ensemble des noeuds, et monitorer l'état des noeuds.
- Automatisation des réactions aux événements.
- Service de fichiers partagés pour la grappe.
- Exemples: OSCAR, OpenPBS, Ganglia
  https://en.wikipedia.org/wiki/Comparison_of_cluster_software

OpenHPC

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Annoncé à SuperComputing 2015, sous l'égide de la Linux Foundation.
- Plusieurs dizaines de membres dont les grands laboratoires gouvernementaux (LANL, LLNL, ANL, BSC, CEA, RIKEN, Calcul Canada...), les fabriquants de puces (Intel, ARM) et d'ordinateurs (Cray, Dell, Fujitsu, HP…) et les distributions Linux (Suse, Red Hat).
- https://github.com/openhpc/ohpc/wiki/Component-List-v1.3.5

Installation d’une grappe

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Noeud de commande installé manuellement. Création d’images pour les noeuds de calcul. Installation par réseau des images avec PXE/DHCP.
- Warewulf, pour installer des grappes de type Beowulf. Ecrit en Perl. http://warewulf.lbl.gov/
- Open Source Cluster Application Resources, OSCAR. Dernière version en 2011, écrit en Perl. https://oscar-cluster.github.io/oscar/
- Beaucoup d’autres outils pour installer des grappes en infonuagique comme MaaS (en Python) sur Ubuntu, et Packstack/ Puppet (en python) sur Red Hat.

Gestion des ressources

- Outil critique pour l'utilisation efficace de la grappe et avec lequel les usagers interagissent directement.
- PBSpro: PBS de la NASA, puis OpenPBS, TORQUE et PBSpro de Altair, écrit en C. Le code de PBSpro a été libéré en 2016 et intégré à OpenHPC.
- SLURM: connait un essor important, support professionnel disponible, license GPLv2, 550 000 lignes de code C.

Conteneurs

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Tous les noeuds d'une grappe ont localement une installation de base et accès aux fichiers des usagers sur un serveur.
- Les tâches sont distribuées sur les noeuds physiques par le gestionnaire de ressources et consistent souvent en un script à rouler.
- Il faut se satisfaire des versions disponibles sur l'installation de base.
- Un conteneur permet d'encapsuler une application dans une image, avec toutes ses dépendances, et de la déployer sur les noeuds.
- Charliecloud construit un conteneur à partir d'une image Docker pour le déployer.
- Singularity permet de construire un conteneur à partir d'une image Docker ou d'un script et de le déployer.

Virtualisation de noeud

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Conteneurs: plusieurs espaces de nom dans le système d'exploitation gérés par le noyau (Linux LXC, Docker).
- Virtualisation logicielle: simulateur d'exécution (Bochs), traducteur dynamique (VMWare, Valgrind).
- Virtualisation matérielle: le processeur peut intercepter ou déléguer certaines instructions privilégiées afin de supporter efficacement la virtualisation (VMWare, KVM).
- Paravirtualisation: le système d'exploitation invité collabore en redirigeant ses requêtes vers le système hôte (Xen, VMWare, KVM).
- Hyperviseur (micro-noyau qui gère les interruptions et protections, e.g. Xen) ou système d'exploitation hôte (e.g. KVM).

Conteneurs

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Un seul noyau avec des espaces de noms séparés par conteneur pour les PID, IPC, usagers, réseau, /proc, hostname, fichiers.
- Chaque conteneur peut rouler une version différente des librairies, applications... mais il n'y a qu'un seul noyau en exécution.
- Aucun coût additionnel en performance, sauf la mémoire non partagée par les versions différentes, si c'est le cas.
- Linux LXC (aussi V-server, OpenVZ), Docker, FreeBSD jails, Solaris containers.

Simulateurs d'exécution

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Programme qui lit et interprète les instructions en simulant le matériel. L'hôte peut être un Intel et l'ordinateur simulé un ARM.
- Certains simulateurs peuvent aussi calculer le nombre de cycles écoulés et s'interfacer à GDB.
- Différentes techniques: interprétation une instruction à la fois, recompilation dynamique par segments, remplacement de certaines instructions et exécution directe des autres.
- BOCHS, QEMU, VMWare et VirtualBox sans support matériel, Valgrind.
- Environ 2 (remplacement de certaines instructions), 5 (recompilation dynamique) ou 50 (interprétation) fois plus lent.

Virtualisation matérielle

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Support matériel (Intel VT, AMD V) pour intercepter ou rediriger certaines opérations.
- Assigner un périphérique à une VM (PCI passthrough), démultiplexer par VM les arrivées de paquets dans la carte réseau, déléguer la table de pages...
- Linux KVM, VMWare et VirtualBox avec support matériel.
- Entre même vitesse et 2 fois plus lent selon le degré d'E/S et d'interaction avec le système d'exploitation.

Virtualisation de table de pages

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Intercepter et émuler toutes les instructions d'accès de la machine virtuelle à la table de pages.
- Détecter et propager à la vraie table de page tous les accès de la machine virtuelle à sa table de pages.
- Intel EPT: *Extended Page Tables* support matériel pour tables de pages dans la machine virtuelle (adresses VM logiques à VM-physiques) en plus de la table de pages usuelle. Pas de coût pour la mise à jour de la table de pages de la machine virtuelle, surcoût pour la traduction si la valeur n'est pas dans le TLB.
- Principale cause de ralentissement dans une machine virtuelle, avec EPT gain de 20% environ (2% ou 3% plus lent si pas de mise à jour, 50% plus rapide pour une application exigeante pour la gestion de mémoire, et 600% plus lent pour un micro-benchmark).

Paravirtualisation

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Le système d'exploitation est modifié pour faire un appel efficace au système d'exploitation hôte. Pas besoin d'intercepter les opérations d'accès au matériel et d'émuler le matériel.
- Plus d'une centaine d'opérations de bas niveau du noyau Linux (lire CR0, désactiver interruptions, lire bloc, changer table de page...) sont appelées à travers la table paravirt_ops qui pointe vers la fonction native ou virtualisée.
- Utilisé par Xen mais aussi VMWare, VirtualBox et KVM avec certains pilotes d'interface virtualisés (disque, réseau, affichage).
- Entre même vitesse et deux fois plus lent, selon le type de charge et les opérations qui sont virtualisées ou non.

Hyperviseur

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Linux virtuel sous Windows réel ou l'inverse?
- Hyperviseur, système d'exploitation minimal qui gère les interruptions et les accès aux périphériques, pour les répartir entre les systèmes d'exploitation des machines virtuelles.
- Xen. Linux domaine 0 qui parle aux périphériques et Linux domaines 1, 2... qui sont les machines virtuelles invitées dont les requêtes sont passées par Xen au domaine 0.
- Xen peut maintenant accepter des invités Windows grâce au support de virtualisation matériel.
- Hyperviseur: solution élégante ou un OS de plus inutilement?

Bénéfices de la virtualisation

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Image logicielle isolée du matériel, utile lorsque les licenses sont attachées au matériel ou pour portabilité.
- Possibilité de cohabitation entre plusieurs systèmes d'exploitation ou versions, plutôt que double amorçage.
- Isolation des services à des fins de sécurité ou de gestion. Plusieurs serveurs virtuels de différents groupes peuvent coexister sur le même serveur physique.
- Modularisation des services: démarrer les serveurs virtuels voulus: base de donnée, courriel, Web...

Coût de la virtualisation

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Certaines instructions causent des interruptions et sont émulées; moins avec le support matériel.
- Accès indirect aux périphériques; moins avec la paravirtualisation ou la virtualisation des I/O (IOMMU).
- Changements de contexte plus nombreux, application, système d'exploitation invité, hyperviseur; moins avec la délégation de tables de pages aux invités.
- Préallocation de la mémoire à chaque machine virtuelle (Xen), n'est pas toujours requis (Xen balloon, KVM).
- Surcoût d'avoir plusieurs copies en mémoire du noyau et des exécutables courants (libc, bash...); moins avec Kernel Samepage Merging.

Virtualisation du réseau

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Réseaux et commutateurs virtuels à l'intérieur d'un noeud pour connecter les noeuds virtuels; Linux TUN/TAP (network tunnel, network tap).
- VLAN: réseau local virtuel séparé du reste du réseau local (étiquette ajoutée à chaque paquet Ethernet, gestion des diffusions générales sur le VLAN).
- VPN/VPLS: connexions multi-point virtuelles privées par-dessus le réseau public.
- Le résultat est un réseau dédié virtuel (overlay network); latence, bande passante, qualité de service... OpenDaylight / OpenFlow

Migration

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Pour équilibrer la charge ou libérer le matériel qui requiert un entretien.
- Déplacer une image en exécution d'une machine virtuelle à l'autre; revient à migrer une machine virtuelle d'un ordinateur physique à un autre de manière transparente.
- Contraintes de même réseau local, mêmes fichiers accessibles, pas de 64 vers 32 bits, matériel virtuel identique.
- Copier toutes les pages de l'image en traçant celles qui sont remodifiées dans l'intervalle. Faire une seconde et possiblement troisième passe. Tout suspendre, copier les pages encore modifiées et poursuivre sur l'autre ordinateur.

Calcul parallèle et virtualisation

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Coût: 100 noeuds + chassis + réseau + aménagements ($500000), ingénieur + technicien + gestionnaire + espace + électricité pendant 5 ans (5 * $300000), total 2M$.
- Location de 100 noeuds pendant 5 ans à 1$/noeud/heure, 4.3M$. Le prix peut descendre mais attention aux frais de stockage, réseau et autres.
- Confidentialité, fiabilité, juridiction.

Discussion

ÉCOLE POLYTECHNIQUE MONTRÉAL

- Les avantages de l'infonuagique peuvent intéresser les utilisateurs de calcul parallèle (flexibilité, capacité à la demande, gestion simplifiée…).
- Une grande grappe pleinement utilisée bénéficie d'une économie d'échelle, sans souffrir du coût d'un intermédiaire, contrairement à une petite compagnie avec des besoins qui varient beaucoup dans le temps pour son site Web et ses rapports.
- Les utilisateurs de calcul parallèle ne veulent pas perdre ~10% de performance pour la virtualisation.
- Serveurs de calcul parallèle sur Amazon, OpenStack avec des conteneurs ou des instances natives "bare metal" et un gestionnaire de ressources…
- Sortir du modèle MPI pour bénéficier de tolérance aux pannes…