Introduction aux systèmes répartis

Grappes de stations

Applications réparties à grande échelle

Systèmes multicalculateurs (1)

- Recherche de puissance par assemblage de calculateurs standard
- Liaison par des réseaux à haut débit
- Grilles = ensemble de grappes éventuellement hétérogènes

Systèmes multicalculateurs (2)

- Ordonnanceur :
  - Coût (en temps) de mémoire partagé important et de changement de noeud(=ordinateur)
    - Un processus est exécuté sur un ordinateur

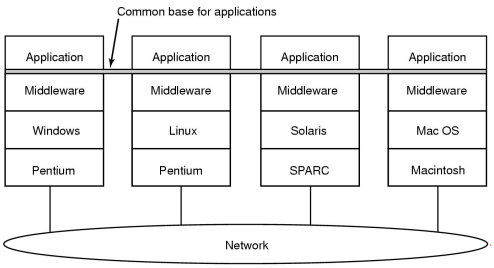
Systèmes multicalculateurs (3)

- Ordonnanceur :
  - Ordonnanceur : Sélectionner le noeud sur lequel sera exécuté le processus
    - Critère : surcharge
      - Au démarrage si surchargé : demander à un autre noeud
      - A l'arrêt d'un processus, demander si processus à exécuter.
      - Offre/demande : vitesse, co-processeur, mémoire …

Systèmes et applications répartis à grande échelle

- Adaptation à la logique des applications
- Répartition de la charge
- Résistance aux pannes
- Une application fait intervenir un ensemble de machines connectées par un réseau
  - Réseau local
  - internet

Intergiciel "middleware"



A la base : l'architecture Client/ Serveur

- Serveur : fournisseur de service.
  - Processus en attente de requête
  - A la réception d'une requête :
    - Traitement
    - Envoi d'une réponse
- Client : consommateur d'un service
  - Emet une requête à un serveur
  - Récupère la réponse à la requête

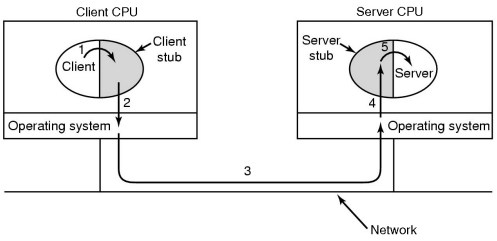
Système utilisant les services réseaux

- Socket
  - UDP
    - Non connecté, non acquité
    - Permet le multicast
  - TCP
    - Connecté, acquité, gestion de flux
    - Socket passive
    - Socket d'échange.

Appel de procédure à distance

- Modèle client-serveur
- Introduction d'objets intermédiaires
  - Les talons (« stubs »)
  - Cachent la répartition des deux côtés
- Prise en compte des pannes
- Annuaire (portmap) des rpc disponibles.

Remote Procedure Call (1)



- Etapes pour exécuter une RPC
  - Les stubs est en gris

Remote Procedure Call (2)

Quelques difficultés

- Impossibilité de passer des pointeurs
  - On passe des valeurs qui doivent être copiées à l'appel et au retour
- Pas de variables globales
- Problèmes de typage
  - Little et big endians

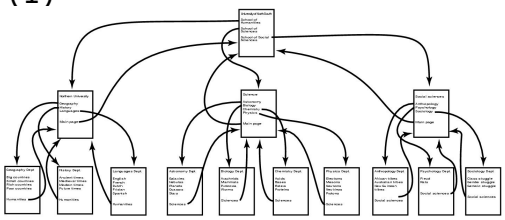
Remote Procedure Call (3)

- Comportement en présence de pannes
- Panne du client, du serveur, du système de communication
- Exécution
  - Au moins une fois
  - Au plus une fois
  - Exactement une fois

Remote Procedure Call (3)

- Comportement en présence de pannes
- Panne du client, du serveur, du système de communication
- Exécution
  - Au moins une fois
  - Au plus une fois
  - Exactement une fois

Document-Based Middleware (1)



- The Web
  - a big directed graph of documents

Document based Middleware(2)

- Comment obtenir une page
  - Demande au DNS d'une adresse IP
  - Réponse du DNS
  - Etablissement d'une connexion au serveur
  - Demande de la page
  - Réception de la page
  - Fin de la connexion
  - Affichage de la page

Navigation web : les points clés

- Représentation uniforme des documents
- Désignation + système de recherche des documents
- Protocole de communication

Désignation dans les systèmes répartis

- Adresse de machine
  - Adresse ethernet, adresse IP
- URL ensimag.grenoble-inp.fr
- Serveur de noms (DNS)
- Désignation d'un objet sur une machine
  - Nom de fichier
  - Nom de processus
  - Port de communication

DNS : Domain Name Server

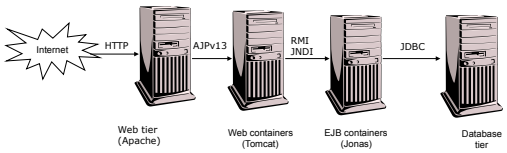
- Espace de nom arborescent muon.inrialpes.fr
- Machine muon, domaines fr et inrialpes.fr
- À chaque domaine est associé une machine : le serveur de noms
  - Catalogue exact de toutes les machines du domaine ainsi que des serveurs de nom des sous-domaines + adresse IP du serveur racine
  - Sur chaque machine, adresse du serveur de nom du domaine le plus bas

Résolution d'un nom

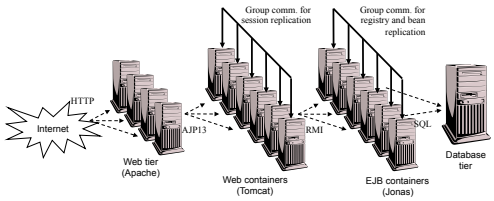
- Localisation de www.cs.unm.edu
- Serveur de noms inrialpes.fr
  - Adresse IP racine qu'on interroge
- Fiabilité : duplication des serveurs
- Efficacité : cache
  - Chaque serveur garde les dernières traductions effectuées

Application multi-niveaux : J2EE architecture

- Serveur de "commerce électronique"



Architecture J2EE sur grappe

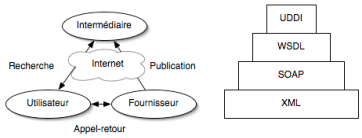


Services web

- Offre de services à des machines clientes
  - Nécessité de contrôle
  - Publication du service : enregistrement auprès d'un intermédiaire
- Client recherche un service
  - Demande à l'intermédiaire
  - Retourne localisation et interface du service, puis appel direct

Les services web

- UDDI : Universal Description Discovery and Integration
- WSDL : Web Service Definition Langage
- SOAP : Simple Object Application Protocol



a) Publication et localisation des services

b) Langages et protocoles

Avantages et inconvénients des systèmes répartis

+ Partage de ressources, de fichiers
+ Puissance de calcul, et souplesse d'augmentation
+ Résistance aux pannes

- Logiciel complexe
- Sensibilité aux communications (panne, saturation)
- Sécurité(intrusions)

Problèmes d'état et de prise de décision

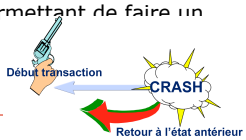
- En centralisé, état représenté par des tables toujours disponibles
- En réparti, problèmes dus aux délais de transmission
- Exemple : producteur consommateur

Transaction

- Atomicité
- Cohérence
- Isolation
- Durabilité

Exemple transaction bancaire

- Compte A donne ordre de virement de X€ vers le compte B
  - X ajouté à B
  - X retiré de A
- Atomicité : Tout faire ou rien
  - Mécanisme permettant de faire un rollback



Validation (commit) en 2 phases

- Processus coordinateur :
  - Fin de transaction, un processus (esclave) exécutant une phase de la transaction envoi terminé au coordinateur
  - Quand tous les esclaves ont terminés :
  - Phase 1 : C->E : prêt_à_valider
    - E->C : OK ou NOK
  - Phase 2 : Si tous OK
    - C-> Es : valider
  - Sinon C->Es : annuler (rollback)

Problèmes d'ordre et d'heure

- Les contraintes
  - Imprécision relative des horloges
  - Grande variabilité des délais de transmission
- Conséquence : il faut mettre de l'ordre
  - Estampilles
  - Anneau virtuel

Estampilles (1)

- Création d'un ordre total strict
- Ordre compatible avec la causalité
- Horloge locale à chaque site $H_i(a)$
- $H_i(a)$ augmentée de 1 à chaque nouvel événement
- L'heure d'émission accompagne chaque message
- A la réception, redéfinition de l'horloge du site d'arrivée si $H_j(b) < H_i(a)$, alors $H_j(b) = H_i(a)+1$

Estampilles (2)

- Soit 2 événements a et b d'estampilles
  $H_i(a)$ et $H_j(b)$
- A précède b ssi $H_i(a) < H_j(b)$ ou $(H_i(a)=H_j(b)$ et $i<j)$

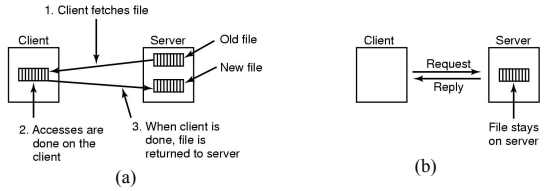
Anneau virtuel

- Tous les sites sont numérotés de 0 à n-1
- Le site i ne communique qu'avec le site i+1(mod n)
- Un message spécial, le jeton, parcourt l'anneau en véhiculant des informations utiles à l'application

Accès à des fichiers distants

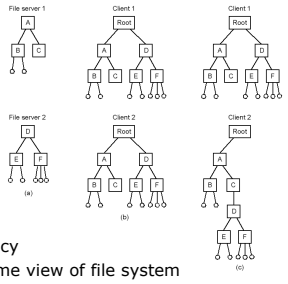
- Généralités
- NFS

File System-Based Middleware (1)



- Transfer Models
  - (a) upload/download model
  - (b) remote access model

File System-Based Middleware (2)



Naming Transparency
(b) Clients have same view of file system
(c) Alternatively, clients with different view

Partage de fichiers : NFS

- Unification du système de désignation
  - Exportation et montage
- Pas de session sur le serveur
- V-node
- Gestion des caches

Quelques points clés pour conclure

- Transparence (localisation, migration, etc.)
- Fiabilité
- Performances
- Passage à l'échelle

Vers des systèmes autonomes

- On ne peut gérer manuellement des systèmes composés de milliers de machines
- Les applications doivent se construire automatiquement
- Elles doivent aussi s'administrer automatiquement
  - Résistance aux pannes
  - Régulation de charge