GRID II

Cuprins – GRID II

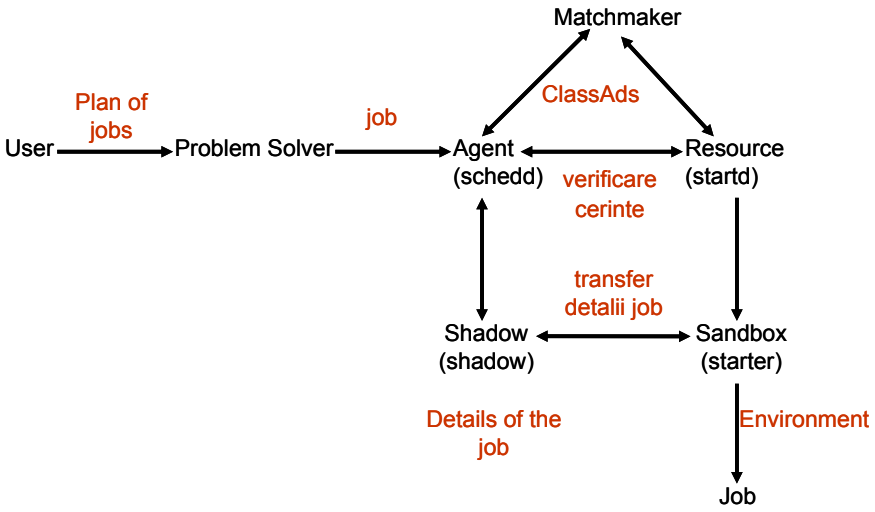
I. Controlul Execuției în Grid

In Grid se face o planificare a activitatilor (taskurilor) la nivelul VO (meta-scheduling) corelata cu o planificare locala,la nivelul resurselor fizice folosite. Ne referim aici la a doua categorie de planificare, pentru care discutam solutiile adoptate in Condor.

Condor este un sistem de “batch queuing” folosit în planificarea job-urilor computational intensive. El reprezinta un mediu *High Throughput Computing* (HTC) care face queueing şi scheduling pentru joburi. Scopurile urmarite sunt folosirea eficientă a resurselor si asigurarea toleranței la defecte. Acestea sunt realizate printr-o abordare oportunistă, constand in:

- Utilizarea resurselor atunci când sunt disponibile
- Job checkpoint, deci salvarea periodica a starii programului pentru a relua executia in cazul unei erori, astfel incat sa se asigure toleranta la defectari, si
migrare de joburi, de exemplu scheduling pre-emptiv pentru a permite intreruperea unui job pentru ca alt job cu prioritate mai mare sa foloseasca resursele puse la dispozitie.

Kernelul Condor permite derularea operatiilor in sisteme bach distribuite si se bazeaza pe cateva componente:



Agent (schedd) este punctul de intrare pentru utilizatori care supun joburi pentru executie

- asigură menținerea persitentă a sarcinilor
- căutarea resurselor de încredere, potrivite pentru job (de ex. memoria necesara)

shadow

- proces creat de agent care asigura detalile legate de executie componentei sandbox (executabile, argumente, fişiere de intrare.....).

Sandbox (starter)

- un mediu sigur de execuție creat pe respectiva resursa, rezistent la interventii exterioare malitioase

startd – serviciul de calcul care gestioneaza o masina si asigura ca executia joburilor se face in contextul restrictiilor impuse de proprietar (ex. joburi de la fizica in timpul zilei, restul noaptea).

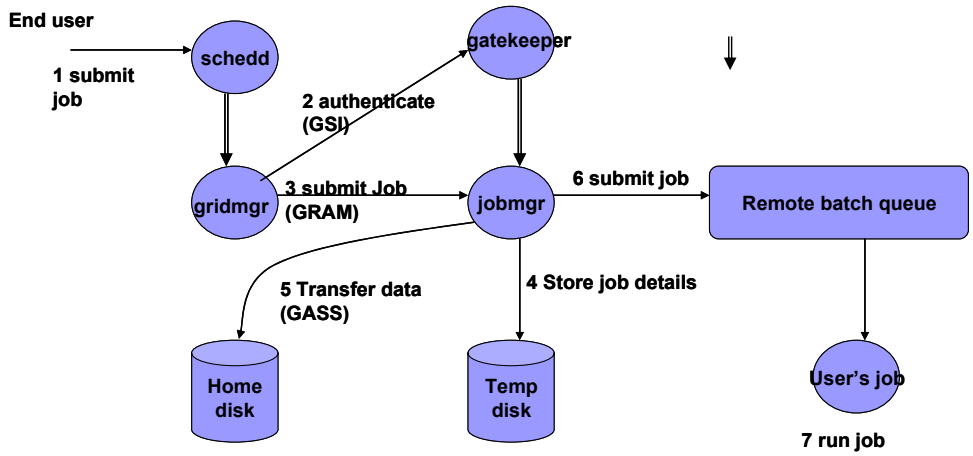
Matchmaker accepta "avertizari" de la consumatori (schedd) si producatori (startd) de resurse, in forma unor descrieri numite ClassAds, si crează perechi de (proces agent, resursa compatibilă), informand ambele parti asupra rezultatului. Se cer autentificare şi autorizare separate.

ClassAds reprezinta perechi (nume atribut – valoare atribut) transmise de agenți şi de resurse la matchmaker, care formează perechi potrivite folosind o logică trivalentă: true, false şi undefined. Pentru potrivire, constrângerile trebuie evaluate la *true*.

Job ClassAd
[
MyType = ‘‘Job’’
TargetType = ‘‘Machine’’
Requirements =
((other.Arch==‘‘INTEL’’&&
other.OpSys==‘‘LINUX’’)
&& other.Disk > my.DiskUsage)
Rank = (Memory ∗ 10000) + KFlops
Cmd = ‘‘/home-exe’’
Department = ‘‘CompSci’’
Owner = ‘‘tannenba’’
DiskUsage = 6000
]

Machine ClassAd
[
MyType=“Machine”
TargetType=“Job”
Machine=“tnt.isi.edu”
Requirements=
(Load<3000)
Rank=dept==self.dept
Arch=“Intel”
OpSys=“Linux”
Disk=600000
]

Rularea unui job in Grid sub Condor-G presupune parcurgerea urmatorilor pasi:

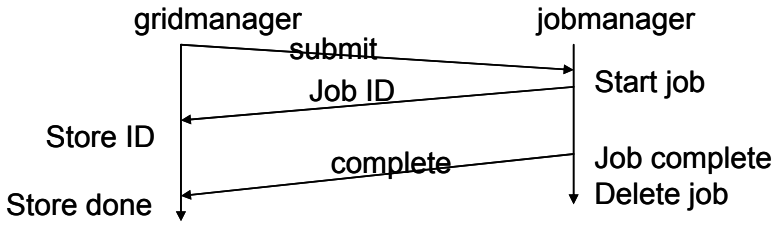


(1) Utilizatorul supune un job pentru executie. Ca urmare, schedd creaza un proces gridmanager pentru acest job
(2) Gridmanager autentifica utiliz la gatekeeper (GSI – Grid Security Infrastructure) care creaza un jobmanager. Jobmanager ofera un mediu de executie la distanta peste un sistem de executie batch existent.

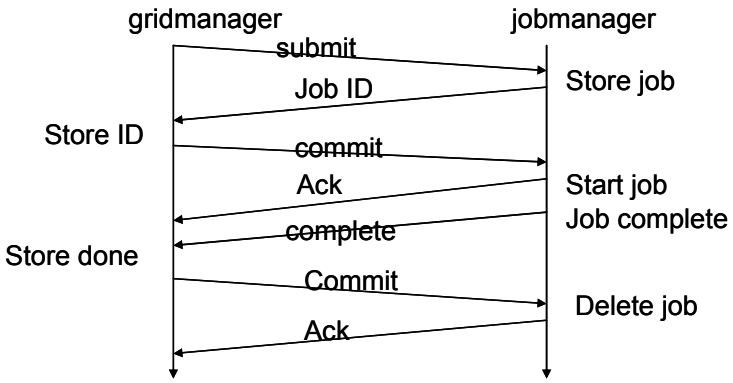
(3) Gridmanager transmite detaliile job-ului catre jobmanager (prin GRAM – Grid Resource Allocation Manager)
(4) Jobmanager memoreaza detaliile
(5) Jobmanager transfera executabilele si datele de intrare de la situl gazda (prin protocolul GASS – Global Access to Secondary Storage cu gridmanager)
(6) Jobmanager supune jobul catre Remote batch queue
(7) Jobul este executat

Globus Resource Allocation Manager (GRAM) realizeaza managementul execuției Job-urilor: pornire, monitorizare, scheduling, coordonare resurse la distanță. Job-urile sunt task-uri computaționale care pot executa operații de I/O si pot modifica starea resursei pe care rulează. Astfel, starea în care se află task-ul se poate modifica pe parcursul execuției acestuia; fişierele resursei pot fi modificate, noi fişiere pot fi create sau aduse de pe alte resurse; incărarea resursei se poate modifica pe parcursul executiei.
Job-urile pot necesita coordonarea aducerii de fişiere pe resursă înainte de execuția task-ului şi mutarea fişierelor rezultate pe alte resurse. Aceasta se poate face si incremental, fiind permis accesul partial la fişierele rezultate, pe măsura ce datele sunt produse în timpul execuției job-ului.
Monitorizarea presupune cererea de şi/sau subscrierea la notificări privind modificările stării job-ului

Protocolul GRAM folosea, in forma originara, interactiuni atomice pentru submiterea si terminarea joburilor. Varianta era pasibila de erori datorate pierderii mesajului "Job ID" (caz in care avem un job orfan – se pierde relatia cu clientul care a declansat jobul), sau pierderii mesajului "complete" (situatie in care jobul este pierdut).



O varianta imbunatatita foloseste "two phase commit", conform schemei date in continuare.



Prima faza include un mesaj de submitere a unui job de catre gridmanager, insotit de date. jobmanager salveaza informatia in memoria permanenta, genereaza un Job ID si il comunica

gridmanager-ului. In a doua faza, gridmanager trimite un mesaj de comitere, confirmat de jobmanager, care demareaza executia jobului. O tehnica similara se foloseste la terminarea jobului.

II. Transferul datelor

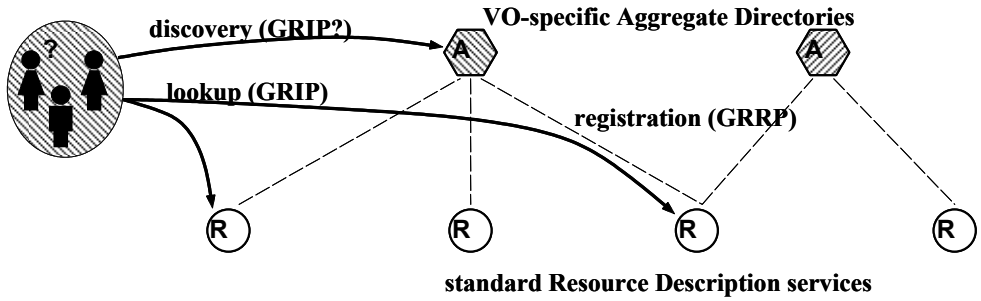
GridFTP - Versiune extinsă a FTP - a fost proiectat ca modalitate fundamentala de acces la date si de transfer de date, pentru o gama larga de sisteme de stocare a datelor. Include failitatile suportate de FTP standard si extensii:
Negociere automată dimensiuni buffer/window – permite setarea manuala sau negocierea automata a dimensiunilor, cu efecte importante asupra performantelor.
Transfer Paralel date: un server transmite in paralel pe mai multe conexiuni TCP.
Transfer controllat de către "**third-party**"; un utilizator poate initia, monitoriza si controla un transfer intre servere de memorare a datelor.
Transfer parțial de fişiere – important in apllicatii in care se prelucreaza subseturi de dimensiuni reduse ale unor fisiere voluminoase.
Securitate: suporta GSI (Grid Security Infrastructure) şi Kerberos furnizand autentificare, confidentialitate si integritate la accesul datelor sau pe durata transferului lor.
Suport pentru **transfer sigur** şi **re-startabil**
Transfer "Striped" sau "intrercalat" pentru fisiere distribuite pe mai multe noduri.

III. Monitoring and Discovery Service

Problemele ce afecteaza activitatea VOs se datoreaza distributiei utilizatorilor si a resurselor. Defectarea unor legaturi sau a unor noduri poate partitiona reteaua si poate conduce la pierderea conectivitatii la nivelul unei VO. Serviciul de monitorizare si descoperire joaca un rol esential in mentinerea continuitatii functionale a Grid-ului in conditiile variabilitatii gruparilor de utilizatori si conectivitatii retelei.

Monitoring and Discovery Service - MDS-2 include doua **servicii de descriere** a serviciilor: unul care permite obtinerea informatiei direct de la fiecare resursa din Grid, al doilea care permite gasirea informatiei in Index-uri ce combină informația de la mai multe surse.

MDS-2 se bazeaza pe doua protocoale:
- Grid Resource Reg. Protocol (GRRP) realizeaza **inregistrarea** dinamică a informatiei oferite de resurse in directoare;
- Grid Resource Info. Protocol (GRIP) permite **descoperirea** resurselor fie prin interogare directa fie prin consultarea directoarelor.



GRRP este un protocol "soft state care porneste de la ideea ca resursele nu pot garanta actualizarea starii lor in indexuri (daca un nod cade, el nu poate avertiza serviciul de directoare despre schimbarea starii). Ca urmare, actualizarea se bazeaza pe diferite mecanisme, cum ar fi:

- Notificare periodică facuta de resurse ca sunt inca disponibile
- Construcție automată a directoarelor fie prin adăugare resursă nouă la un director sau prin invitarea resurselor să se alăture la un director
- "Self-cleaning" care permite reducerea referințelor "moarte"

Informațiile sunt publicate conform unei scheme, de exemplu:
- MDS-2 foloseşte LDAP
- MDS-3 foloseşte XML

MDS poate fi privit si ca o suită de servicii web pentru monitorizare şi descoperirea de resurse şi servicii Grid, care permite Descoperirea de resurse ale VO si are interfețe pentru interogarea şi subscrierea la informații privind resursele servicilor Grid.

MDS4 utilizeaza trei metode pentru a obține date cu privire la resurse
- Interogarea serviciilor Grid care utilizează WSRF pentru informații privind proprietățile resurselor (WS ResourceProperties)
- Executarea de programe generale pentru preluarea de date
- Interfețe la sisteme de monitorizare *third-party*.

MDS4 este o suită de servicii de monitorizare oferită de GT4.0
MDS-Index – cu care se inregistreaza serviciile - adună info de la mai multe servere si suportă căutare după caracteristici.
Ex1: ce maşini au >16 procesoare?
Ex2: ce servere de memorie au bandă >100Mbps la gazda X?
MDS-Trigger – declanseaza o actiune cand datele stranse indeplinesc anumite conditii.

Referințe

Ian Foster, Carl Kesselman, Steven Tuecke – The Anatomy of the Grid Enabling Scalable Virtual Organizations
I. Foster, C. Kesselman, J. Nick, S. Tuecke, - The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration.
Foster. - Globus Toolkit Version 4: Software for Service-Oriented Systems.
I. Foster (ed), J. Frey (ed), S. Graham (ed), S. Tuecke (ed), K. Czajkowski, D. Ferguson, F. Leymann, M. Nally, I. Sedukhin, D. Snelling, T. Storey, W. Vambenepe, S. Weerawarana, Modeling Stateful Resources with Web Services v. 1.1.
Globus Toolkit documentation pages - http://www.globus.org/toolkit/docs/4.0/
Ian Foster - A Globus Primer Or, Everything You Wanted to Know about Globus, but Were Afraid To Ask
Borja Sotomayor - http://gdp.globus.org/gt4-tutorial/singlehtml/progtutorial_0.2.html
V. Welch, I. Foster, C. Kesselman, O. Mulmo, L. Pearlman, S. Tuecke, J. Gawor, S. Meder, F. Siebenlist - X.509 Proxy Certificates for Dynamic Delegation.
V. Welch, F. Siebenlist, I. Foster, J. Bresnahan, K. Czajkowski, J. Gawor, C. Kesselman, S. Meder, L. Pearlman, S. Tuecke – Security for Grid Services.
Ian Foster, Carl Kesselman, Laura Pearlman, Steven Tuecke, and Von Welch - The Community Authorization Service: Status and Future.
J. Novotny, S. Tuecke, V. Welch - An Online Credential Repository for the Grid: MyProxy.
Globus Toolkit presentations: http://www.globus.org/toolkit/presentations/