Exploitation d'un calculateur

Monitoring de clusters, pourquoi, comment ?

Olivier Brand-Foissac

CNRS / LPT / ARGOS-Rodia / RESINFO

RESINFO - ANGD - 9 octobre 2009

Monitoring de clusters

Plan

Monitoring de clusters

Plan

Définitions
monitoring et supervision

Monitoring

Monitoring means the periodic inspection by [..] a directed function or activity and includes watching during performance, checking, and tracking progress, updating a supervisor of progress or accomplishment by the person monitored, and contacting a supervisor as needed for direction and consultation.

Supervision

"Supervision" means the guidance by a registered one for the accomplishment of a function or activity. The guidance consists of the activities included in monitoring as well as establishing the initial direction, delegating, setting expectations, directing activities and courses of action, critical watching, overseeing, evaluating, and changing a course of action.

Source [Minnesota Administrative Rules - 2008]

Définitions
monitoring et supervision

Monitoring Le *monitoring* signifie l'inspection périodique par une fonction dirigée ou une activité et inclue la surveillance pendant l'action, le contrôle, le suivi de process, informant un superviseur des progrès ou accomplissement par l'objet du monitoring, et contactant un superviseur au besoin pour des directions et des consultations.

Supervision La *supervision* signifie l'orientation en vue de l'accomplissement d'une fonction ou d'une activité inclue dans le *monitoring* autant que dans l'établissement des directions initiales, des délégations, de la définition des attendus, dirigeant les activités et les plans d'action, un regard critique, surveillant, évaluant et modifiant un plan d'action.

Définitions
monitoring et supervision

Monitoring Le *monitoring* signifie l'inspection périodique par une fonction dirigée ou une activité et inclue la surveillance pendant l'action, le contrôle, le suivi de process, informant un superviseur des progrès ou accomplissement par l'objet du monitoring, et contactant un superviseur au besoin pour des directions et des consultations.

Supervision La *supervision* signifie l'orientation en vue de l'accomplissement d'une fonction ou d'une activité inclue dans le *monitoring* autant que dans l'établissement des directions initiales, des délégations, de la définition des attendus, dirigeant les activités et les plans d'action, un regard critique, surveillant, évaluant et modifiant un plan d'action.

Définitions
monitoring et supervision

En informatique, *monitoring* et *supervision* se distinguent par :

Monitoring
- local, courte portée, proximité
- précision
- temps réel
- orienté diagnostic
- attaché à la performance

Supervision
- global, longue portée, regrouppement
- consolidation, concaténation, concentration
- temps différé
- orienté présentation, pilotage
- attaché au service

Définitions
monitoring et supervision

En informatique, *monitoring* et *supervision* se distinguent par :

Monitoring
- local, courte portée, proximité
- précision
- temps réel
- orienté diagnostic
- attaché à la performance

Supervision
- global, longue portée, regrouppement
- consolidation, concaténation, concentration
- temps différé
- orienté présentation, pilotage
- attaché au service

Définitions
conclusion

Les domaines *monitoring* et *supervision* ne sont pas disjoints.

Beaucoup confondent les deux aspects, les outils débordant souvent de l'un sur l'autre.

La supervision nécessite le monitoring et constitue son aboutissement.

Limites
portée du cours

Cantonnement :

- outils gratuits (et libres)
- principalement sur systèmes GNU-Linux
- processeurs généralistes (Intel/AMD - like)
- autres hardware : les principes restent vrais,
  faire préciser par les constructeurs les accès aux autres caractéristiques (compteurs hard, librairies de diagnostique, etc.)

Non abordé :

- la gestion des clusters
- l'administration (choix systèmes, compilteurs, benchmarks, . . .)
- l'exhaustivité : plus de 1500 projet uniquement sur sourceforge

Les principes présentés sont en général transposables.

Limites
portée du cours

Cantonnement :

- outils gratuits (et libres)
- principalement sur systèmes GNU-Linux
- processeurs généralistes (Intel/AMD - like)
- autres hardware : les principes restent vrais,
  faire préciser par les constructeurs les accès aux autres caractéristiques (compteurs hard, librairies de diagnostique, etc.)

Non abordé :

- la gestion des clusters
- l'administration (choix systèmes, compilteurs, benchmarks, . . .)
- l'exhaustivité : plus de 1500 projet uniquement sur sourceforge

Les principes présentés sont en général transposables.

Limites
portée du cours

Cantonnement :

- outils gratuits (et libres)
- principalement sur systèmes GNU-Linux
- processeurs généralistes (Intel/AMD - like)
- autres hardware : les principes restent vrais,
  faire préciser par les constructeurs les accès aux autres caractéristiques (compteurs hard, librairies de diagnostique, etc.)

Non abordé :

- la gestion des clusters
- l'administration (choix systèmes, compilteurs, benchmarks, . . .)
- l'exhaustivité : plus de 1500 projet uniquement sur sourceforge

Les principes présentés sont en général transposables.

Une nécessité ?
sans aucun doute

En quoi le monitoring est-il nécessaire ?

- contrôlee la disponibilité des services/fonctions
- contrôler l'utilisation des ressources
- contrôler qu'elles sont suffisantes
- vérifier l'équilibrage de charge
- diagnostic de panne (pannes avérées)
- prévention des pannes/défauts (pannes latentes)
- prévoir les évolutions (gestion de cluster)
  - en terme de ressources (CPU, stockage, fluides, . . .)
  - en terme de capacités (accès et utilisation des ressources)
  - en terme de disponibilités (attente en files, contentions, HA)

Une nécessité ?

sans aucun doute

En quoi le monitoring est-il nécessaire ?

- contrôlee la disponibilité des services/fonctions
- contrôler l'utilisation des ressources
- contrôler qu'elles sont suffisantes
- vérifier l'équilibrage de charge
- diagnostic de panne (pannes avérées)
- prévention des pannes/défauts (pannes latentes)
- prévoir les évolutions (gestion de cluster)
  - en terme de ressources (CPU, stockage, fluides, . . .)
  - en terme de capacités (accès et utilisation des ressources)
  - en terme de disponibilités (attente en files, contentions, HA)

Une nécessité ?

sans aucun doute

En quoi le monitoring est-il nécessaire ?

- contrôlee la disponibilité des services/fonctions
- contrôler l'utilisation des ressources
- contrôler qu'elles sont suffisantes
- vérifier l'équilibrage de charge
- diagnostic de panne (pannes avérées)
- prévention des pannes/défauts (pannes latentes)
- prévoir les évolutions (gestion de cluster)
  - en terme de ressources (CPU, stockage, fluides, . . .)
  - en terme de capacités (accès et utilisation des ressources)
  - en terme de disponibilités (attente en files, contentions, HA)

Une nécessité ?

sans aucun doute

En quoi le monitoring est-il nécessaire ?

- contrôlee la disponibilité des services/fonctions
- contrôler l'utilisation des ressources
- contrôler qu'elles sont suffisantes
- vérifier l'équilibrage de charge
- diagnostic de panne (pannes avérées)
- prévention des pannes/défauts (pannes latentes)
- prévoir les évolutions (gestion de cluster)
  - en terme de ressources (CPU, stockage, fluides, . . .)
  - en terme de capacités (accès et utilisation des ressources)
  - en terme de disponibilités (attente en files, contentions, HA)

Une nécessité ?

sans aucun doute

En quoi le monitoring est-il nécessaire ?

- contrôlee la disponibilité des services/fonctions
- contrôler l'utilisation des ressources
- contrôler qu'elles sont suffisantes
- vérifier l'équilibrage de charge
- diagnostic de panne (pannes avérées)
- prévention des pannes/défauts (pannes latentes)
- prévoir les évolutions (gestion de cluster)
  - en terme de ressources (CPU, stockage, fluides, . . .)
  - en terme de capacités (accès et utilisation des ressources)
  - en terme de disponibilités (attente en files, contentions, HA)

Une nécessité ?

sans aucun doute

En quoi le monitoring est-il nécessaire ?

- contrôlee la disponibilité des services/fonctions
- contrôler l'utilisation des ressources
- contrôler qu'elles sont suffisantes
- vérifier l'équilibrage de charge
- diagnostic de panne (pannes avérées)
- prévention des pannes/défauts (pannes latentes)
- prévoir les évolutions (gestion de cluster)
  - en terme de ressources (CPU, stockage, fluides, . . .)
  - en terme de capacités (accès et utilisation des ressources)
  - en terme de disponibilités (attente en files, contentions, HA)

Une nécessité ?

sans aucun doute

En quoi le monitoring est-il nécessaire ?

- contrôlee la disponibilité des services/fonctions
- contrôler l'utilisation des ressources
- contrôler qu'elles sont suffisantes
- vérifier l'équilibrage de charge
- diagnostic de panne (pannes avérées)
- prévention des pannes/défauts (pannes latentes)
- prévoir les évolutions (gestion de cluster)
  - en terme de ressources (CPU, stockage, fluides, . . .)
  - en terme de capacités (accès et utilisation des ressources)
  - en terme de disponibilités (attente en files, contentions, HA)

Une nécessité ?

sans aucun doute

En quoi le monitoring est-il nécessaire ?

- contrôlee la disponibilité des services/fonctions
- contrôler l'utilisation des ressources
- contrôler qu'elles sont suffisantes
- vérifier l'équilibrage de charge
- diagnostic de panne (pannes avérées)
- prévention des pannes/défauts (pannes latentes)
- prévoir les évolutions (gestion de cluster)
  - en terme de ressources (CPU, stockage, fluides, ...)
  - en terme de capacités (accès et utilisation des ressources)
  - en terme de disponibilités (attente en files, contentions, HA)

Monitoring de clusters

Plan

Principes généraux

Constituants

Quatre constituants de la chaîne monitoring-supervision :

1. Collecte des données (acquisition).
   - ciblage (ce qui sera mesuré)
   - acquisition (faire la mesure)
     - comment faire la mesure
     - où la faire (actif, passif)
   - stockage (où placer les mesures)
2. Analyse des données recueillies.
   - immédiate
   - en différé

Principes généraux

Constituants

Quatre constituants de la chaîne monitoring-supervision :

1. Collecte des données (acquisition).
   - ciblage (ce qui sera mesuré)
   - acquisition (faire la mesure)
     - comment faire la mesure
     - où la faire (actif, passif)
   - stockage (où placer les mesures)
2. Analyse des données recueillies.
   - immédiate
   - en différé

Principes généraux

Constituants (suite)

Quatre constituants (suite) :

3. Action déclenchée par l'analyse.
   - visualisation graphique (passif, pré-conditionnement)
   - alertes (actif)
   - ré-actions (actif)

4. Pilotage (ou contrôle par l'opérateur).
   - déclencher l'analyse (différée)
   - renouveler mesure/analyse
   - action sur l'objet de la mesure ou sur le système (ouverture/fermeture de ports réseaux, etc.)

Principes généraux

Constituants (suite)

Quatre constituants (suite) :

3. Action déclenchée par l'analyse.
   - visualisation graphique (passif, pré-conditionnement)
   - alertes (actif)
   - ré-actions (actif)

4. Pilotage (ou contrôle par l'opérateur).
   - déclencher l'analyse (différée)
   - renouveler mesure/analyse
   - action sur l'objet de la mesure ou sur le système (ouverture/fermeture de ports réseaux, etc.)

Principes généraux

Acquisition

- Sélection des objets des mesures, les quantités associées (taux de charge, valeur/taux remplissage, présence/absence, etc.)
- Choix des fréquences de collecte
  - granularité et précision des données
  - influence sur le volume
- Choix des outils de collecte
  - avec agent
    (process permanent local (démon), mesures en continu ou temporisées, influence sur les performances de l'hôte)
  - sans agent
    (déclenchement à distance ou local (cron))
- Stockage et format des données
  (local et/ou distant, historique, cumulatif, ...)
  date et lieu

Principes généraux

Acquisition

- Sélection des objets des mesures, les quantités associées (taux de charge, valeur/taux remplissage, présence/absence, etc.)
- Choix des fréquences de collecte
  - granularité et précision des données
  - influence sur le volume
- Choix des outils de collecte
  - avec agent
    (process permanent local (démon), mesures en continu ou temporisées, influence sur les performances de l'hôte)
  - sans agent
    (déclenchement à distance ou local (cron))
- Stockage et format des données
  (local et/ou distant, historique, cumulatif, ...)
  date et lieu

Principes généraux

Acquisition

- Sélection des objets des mesures, les quantités associées (taux de charge, valeur/taux remplissage, présence/absence, etc.)
- Choix des fréquences de collecte
  - granularité et précision des données
  - influence sur le volume
- Choix des outils de collecte
  - avec agent
    (process permanent local (démon), mesures en continu ou temporisées, influence sur les performances de l'hôte)
  - sans agent
    (déclenchement à distance ou local (cron))
- Stockage et format des données
  (local et/ou distant, historique, cumulatif, ...)
  date et lieu

Principes généraux

Acquisition

- Sélection des objets des mesures, les quantités associées (taux de charge, valeur/taux remplissage, présence/absence, etc.)
- Choix des fréquences de collecte
  - granularité et précision des données
  - influence sur le volume
- Choix des outils de collecte
  - avec agent
    (process permanent local (démon), mesures en continu ou temporisées, influence sur les performances de l'hôte)
  - sans agent
    (déclenchement à distance ou local (cron))
- Stockage et format des données
  (local et/ou distant, historique, cumulatif, ...)
  date et lieu

Principes généraux

Analyse

Extraire les informations utiles et exploiter les données recueillies (comparaison de valeurs seuils, recherche de mot-clés, calculs de différences, . . .)

- à destination de concaténation, de regroupement, de filtrage
- à destination d'action (déclenchement d'alertes, . . .)
- à destination visuelle (éléments de tableau de bord, graphiques)
- pré-conditionnement (reformatage avec ou sans perte)

Principes généraux

Actions Informatives

Dans le but de diffuser l'information, selon une criticité établie

- alerte par eMail (listes ciblées, différents niveaux)
- alerte par action locale (trap d'agent, création de fichier, etc.)
- alerte par action distante (dépôt de fichier, télé-alarme, page web)
- garder les traces d'alertes

Principes généraux

Actions Opératives

Dans le but de provoquer des modifications

- Auto-action (pare-feu, . . .)
- Auto-extinction (seuil de température, . . .)
- Purges de données
- Nettoyages divers (caches, historisation, fichiers temporaires, . . .)
- Relance d'acquisition

Principes généraux

Contrôles

Contrôles par l'opérateur :

- directs (visuels, tableaux de bord)
- indirects (recherche dans les traces)
- par consultation des historiques (fréquence de défauts, . . .)

Attention

- Confidentialité (accès)
- Législation

Briques de base

Protocoles

Les briques de base du monitoring :

- librairies (accès aux objets, matériels)
- HTTP/HTTPS
- SMTP
- TCP/IP (multicast)
- SNMP

Briques de base

Protocoles

Simple Network Management Protocol

Outils très répandu

- Agent
  - démon local (snmpd)
  - MIB (Management Information Base)
    MIB-II standard, MIB spécifiques
  - ASN-1 (Abstract Syntax Notation-1)
  - OID (Object IDentifier)
    .1.3.6.1.2.1 (racine de mib-2) = ".iso.org.dod.internet.mgmt.mib"
    .1.3.6.1.2.1.1.1.0 .iso.org.dod.internet.mgmt.mib.system.sysdescr.0 .1.3.6.1.2.1.1.sysdescr.0

    1.1.0 system.sysDescr.0 1.sysDescr.0 sont équivalents (supposés dans la mib-2)
  - adaptable (MIB, scripts, . . .)
- Client
  - consultation ou écriture
  - élément par élément ou bulk
  - Attention : plus de 4500 objets dans la MIB-II

Briques de base

Protocoles

Simple Network Management Protocol
Outils très répandu

- Agent
  - démon local (snmpd)
  - MIB (Management Information Base)
    MIB-II standard, MIB spécifiques
  - ASN-1 (Abstract Syntax Notation-1)
  - OID (Object IDentifier)
    .1.3.6.1.2.1 (racine de mib-2) = ".iso.org.dod.internet.mgmt.mib"
    .1.3.6.1.2.1.1.1.0 .iso.org.dod.internet.mgmt.mib.system.sysdescr.0 .1.3.6.1.2.1.1.sysdescr.0

    1.1.0 system.sysDescr.0 1.sysDescr.0 sont équivalents (supposés dans la mib-2)
  - adaptable (MIB, scripts, . . .)
- Client
  - consultation ou écriture
  - élément par élément ou bulk
  - Attention : plus de 4500 objets dans la MIB-II

Briques de base

Protocoles

Example

snmpwalk -Os -v2c -c *mycommunity mymachine* system.sysDescr
sysDescr.0 = STRING : Linux idefix 2.6.5-7.276-default #1 Mon Jul 24 10 :45 :31 UTC 2006 i686

snmpwalk -v2c -c comsulptse idefix system.sysDescr
SNMPv2-MIB : :sysDescr.0 = STRING : Linux idefix 2.6.5-7.276-default #1 Mon Jul 24 10 :45 :31 UTC 2006 i686

Attention : en terme de sécurité

- Pas (peu) de sécurité avant v3
- Changer les accès par défaut (écriture ?)
- Réseau (VLAN) indépendant

Briques de base

Outils de base

Stockage des mesures :

- fichiers de logs (locaux, centralisés, *syslog-ng*)
- stockages en RRD (Round-Robin Database)
  - valeurs, moyennes, maxi, mini
  - auto-lissée sur le temps
  - taille fixée à la création
- formats spécifique des outils

Principes généraux

Présentation

Interfaces de présentation des résultats

- présentation synthétique (web)
- plus ou moins de détails
- historisation et consolidation statistique
- textes et/ou graphiques (RRD tools)

Principes généraux

Présentation

Interfaces de présentation des résultats

- présentation synthétique (web)
- plus ou moins de détails
- historisation et consolidation statistique
- textes et/ou graphiques (RRD tools)

Principes généraux

Présentation

Interfaces de présentation des résultats

- présentation synthétique (web)
- plus ou moins de détails
- historisation et consolidation statistique
- textes et/ou graphiques (RRD tools)

Principes généraux

Présentation

Interfaces de présentation des résultats

- présentation synthétique (web)
- plus ou moins de détails
- historisation et consolidation statistique
- textes et/ou graphiques (RRD tools)

Monitoring de clusters

Plan

Choix des types

Mesures actives / passives

Sondes actives

Avantages	Inconvénients
✓ proximité	• influence sur l'hôte
✓ pas d'influences extérieures	• stockage
✓ grain fin	
✓ fréquence	

Sondes passives

Avantages	Inconvénients
✓ non intrusives	• limitées aux accès extérieurs
✓ alerte pas absence	• sécurité
✓ volume pour données	• fréquence limitée

Choix des types

Mesures actives / passives

Sondes actives

Avantages	Inconvénients
✓ proximité	• influence sur l'hôte
✓ pas d'influences extérieures	• stockage
✓ grain fin	
✓ fréquence	

Sondes passives

Avantages	Inconvénients
✓ non intrusives	• limitées aux accès extérieurs
✓ alerte pas absence	• sécurité
✓ volume pour données	• fréquence limitée

Choix des types

Stockage des mesures

Stockage local

Avantages	Inconvénients
✓ immédiat	• influence sur l'hôte
✓ fréquences	• accès sur crash
	• volume limité
	• capacité d'alertes limitées

Stockage distant

Avantages	Inconvénients
✓ disponible	• blocage sur isolation
✓ alertes	• sécurité
	• fréquence limitée
	• influence sur réseaux

Choix des types

Stockage des mesures

Stockage local

Avantages	Inconvénients
✓ immédiat	• influence sur l'hôte
✓ fréquences	• accès sur crash
	• volume limité
	• capacité d'alertes limitées

Stockage distant

Avantages	Inconvénients
✓ disponible	• blocage sur isolation
✓ alertes	• sécurité
	• fréquence limitée
	• influence sur réseaux

Choix des types

Précision et performance

Choisir entre précision des mesures et performance de la machine :

Précision :

- fréquence élevée
- sonde embarquée
- impacte CPU et disque de l'hôte

Performance :

- fréquence étagée
- sonde distante
- impacte le réseau

L'idéal dépend souvent de la situation : routine ou crise.

Choix des types

Précision et performance

Choisir entre précision des mesures et performance de la machine :

Précision :

- fréquence élevée
- sonde embarquée
- impacte CPU et disque de l'hôte

Performance :

- fréquence étagée
- sonde distante
- impacte le réseau

L'idéal dépend souvent de la situation : routine ou crise.

Choix des types

Précision et performance

Choisir entre précision des mesures et performance de la machine :

Précision :

- fréquence élevée
- sonde embarquée
- impacte CPU et disque de l'hôte

Performance :

- fréquence étagée
- sonde distante
- impacte le réseau

L'idéal dépend souvent de la situation : routine ou crise.

Choix des types

Précision et performance

Choisir entre précision des mesures et performance de la machine :

Précision :

- fréquence élevée
- sonde embarquée
- impacte CPU et disque de l'hôte

Performance :

- fréquence étagée
- sonde distante
- impacte le réseau

L'idéal dépend souvent de la situation : routine ou crise.

Choix des types

Précision et performance

Choisir entre précision des mesures et performance de la machine :

Précision :

- fréquence élevée
- sonde embarquée
- impacte CPU et disque de l'hôte

Performance :

- fréquence étagée
- sonde distante
- impacte le réseau

L'idéal dépend souvent de la situation : routine ou crise.

Monitoring de clusters

Plan

Spécificités

Que contrôler ?

Au niveau des machines (nœuds de calcul et serveurs)

- les CPU (% sys, % user, % idle, nombre de cores utilisés)
- utilisation de la mémoire (cache, swap, fautes)
- nombre de processus (contentions)
- utilisation des disques (lectures/écritures, wait sur I/O, remplissage, pannes)
- utilisation des réseaux (débits, latences, bande passante utilisée)
- température processeurs, température du boîtier
- vitesse de rotation des ventillateurs
- sécurité (authentifications, tunnels)
- disponibilité des services (batch, interfaces)
- compteurs hardware si disponibles et pertinents (mcelog, . . .)
- sabordage de runs (limite de temps, plantages, bouclages)
- . . .

Spécificités

Que contrôler ?

Au niveau des machines (nœuds de calcul et serveurs)

- les CPU (% sys, % user, % idle, nombre de cores utilisés)
- utilisation de la mémoire (cache, swap, fautes)
- nombre de processus (contentions)
- utilisation des disques (lectures/écritures, wait sur I/O, remplissage, pannes)
- utilisation des réseaux (débits, latences, bande passante utilisée)
- température processeurs, température du boîtier
- vitesse de rotation des ventillateurs
- sécurité (authentifications, tunnels)
- disponibilité des services (batch, interfaces)
- compteurs hardware si disponibles et pertinents (mcelog, . . .)
- sabordage de runs (limite de temps, plantages, bouclages)
- . . .

Spécificités

Que contrôler ?

Au niveau des machines (nœuds de calcul et serveurs)

- les CPU (% sys, % user, % idle, nombre de cores utilisés)
- utilisation de la mémoire (cache, swap, fautes)
- nombre de processus (contentions)
- utilisation des disques (lectures/écritures, wait sur I/O, remplissage, pannes)
- utilisation des réseaux (débits, latences, bande passante utilisée)
- température processeurs, température du boîtier
- vitesse de rotation des ventillateurs
- sécurité (authentifications, tunnels)
- disponibilité des services (batch, interfaces)
- compteurs hardware si disponibles et pertinents (mcelog, . . .)
- sabordage de runs (limite de temps, plantages, bouclages)
- . . .

Spécificités

Que contrôler ?

Au niveau des machines (nœuds de calcul et serveurs)

- les CPU (% sys, % user, % idle, nombre de cores utilisés)
- utilisation de la mémoire (cache, swap, fautes)
- nombre de processus (contentions)
- utilisation des disques (lectures/écritures, wait sur I/O, remplissage, pannes)
- utilisation des réseaux (débits, latences, bande passante utilisée)
- température processeurs, température du boîtier
- vitesse de rotation des ventillateurs
- sécurité (authentifications, tunnels)
- disponibilité des services (batch, interfaces)
- compteurs hardware si disponibles et pertinents (mcelog, . . .)
- sabordage de runs (limite de temps, plantages, bouclages)
- . . .

Spécificités

Que contrôler ?

Au niveau des machines (nœuds de calcul et serveurs)

- les CPU (% sys, % user, % idle, nombre de cores utilisés)
- utilisation de la mémoire (cache, swap, fautes)
- nombre de processus (contentions)
- utilisation des disques (lectures/écritures, wait sur I/O, remplissage, pannes)
- utilisation des réseaux (débits, latences, bande passante utilisée)
- température processeurs, température du boîtier
- vitesse de rotation des ventillateurs
- sécurité (authentifications, tunnels)
- disponibilité des services (batch, interfaces)
- compteurs hardware si disponibles et pertinents (mcelog, . . .)
- sabordage de runs (limite de temps, plantages, bouclages)
- . . .

Spécificités

Que contrôler ?

Au niveau des machines (nœuds de calcul et serveurs)

- les CPU (% sys, % user, % idle, nombre de cores utilisés)
- utilisation de la mémoire (cache, swap, fautes)
- nombre de processus (contentions)
- utilisation des disques (lectures/écritures, wait sur I/O, remplissage, pannes)
- utilisation des réseaux (débits, latences, bande passante utilisée)
- température processeurs, température du boîtier
- vitesse de rotation des ventillateurs
- sécurité (authentifications, tunnels)
- disponibilité des services (batch, interfaces)
- compteurs hardware si disponibles et pertinents (mcelog, . . .)
- sabordage de runs (limite de temps, plantages, bouclages)
- . . .

Spécificités

Que contrôler ?

Au niveau des machines (nœuds de calcul et serveurs)

- les CPU (% sys, % user, % idle, nombre de cores utilisés)
- utilisation de la mémoire (cache, swap, fautes)
- nombre de processus (contentions)
- utilisation des disques (lectures/écritures, wait sur I/O, remplissage, pannes)
- utilisation des réseaux (débits, latences, bande passante utilisée)
- température processeurs, température du boîtier
- vitesse de rotation des ventillateurs
- sécurité (authentifications, tunnels)
- disponibilité des services (batch, interfaces)
- compteurs hardware si disponibles et pertinents (mcelog, . . .)
- sabordage de runs (limite de temps, plantages, bouclages)
- . . .

Spécificités

Que contrôler ?

Au niveau des machines (nœuds de calcul et serveurs)

- les CPU (% sys, % user, % idle, nombre de cores utilisés)
- utilisation de la mémoire (cache, swap, fautes)
- nombre de processus (contentions)
- utilisation des disques (lectures/écritures, wait sur I/O, remplissage, pannes)
- utilisation des réseaux (débits, latences, bande passante utilisée)
- température processeurs, température du boîtier
- vitesse de rotation des ventillateurs
- sécurité (authentifications, tunnels)
- disponibilité des services (batch, interfaces)
- compteurs hardware si disponibles et pertinents (mcelog, . . .)
- sabordage de runs (limite de temps, plantages, bouclages)
- . . .

Spécificités

Que contrôler ?

Au niveau des machines (nœuds de calcul et serveurs)

- les CPU (% sys, % user, % idle, nombre de cores utilisés)
- utilisation de la mémoire (cache, swap, fautes)
- nombre de processus (contentions)
- utilisation des disques (lectures/écritures, wait sur I/O, remplissage, pannes)
- utilisation des réseaux (débits, latences, bande passante utilisée)
- température processeurs, température du boîtier
- vitesse de rotation des ventillateurs
- sécurité (authentifications, tunnels)
- disponibilité des services (batch, interfaces)
- compteurs hardware si disponibles et pertinents (mcelog, . . .)
- sabordage de runs (limite de temps, plantages, bouclages)
- . . .

Spécificités

Que contrôler ?

Au niveau des machines (nœuds de calcul et serveurs)

- les CPU (% sys, % user, % idle, nombre de cores utilisés)
- utilisation de la mémoire (cache, swap, fautes)
- nombre de processus (contentions)
- utilisation des disques (lectures/écritures, wait sur I/O, remplissage, pannes)
- utilisation des réseaux (débits, latences, bande passante utilisée)
- température processeurs, température du boîtier
- vitesse de rotation des ventillateurs
- sécurité (authentifications, tunnels)
- disponibilité des services (batch, interfaces)
- compteurs hardware si disponibles et pertinents (mcelog, . . .)
- sabordage de runs (limite de temps, plantages, bouclages)
- . . .

Spécificités

Que contrôler ?

Au niveau des machines (nœuds de calcul et serveurs)

- les CPU (% sys, % user, % idle, nombre de cores utilisés)
- utilisation de la mémoire (cache, swap, fautes)
- nombre de processus (contentions)
- utilisation des disques (lectures/écritures, wait sur I/O, remplissage, pannes)
- utilisation des réseaux (débits, latences, bande passante utilisée)
- température processeurs, température du boîtier
- vitesse de rotation des ventillateurs
- sécurité (authentifications, tunnels)
- disponibilité des services (batch, interfaces)
- compteurs hardware si disponibles et pertinents (mcelog, . . .)
- sabordage de runs (limite de temps, plantages, bouclages)
- . . .

Spécificités

Que contrôler ?

Au niveau des machines (nœuds de calcul et serveurs)

- les CPU (% sys, % user, % idle, nombre de cores utilisés)
- utilisation de la mémoire (cache, swap, fautes)
- nombre de processus (contentions)
- utilisation des disques (lectures/écritures, wait sur I/O, remplissage, pannes)
- utilisation des réseaux (débits, latences, bande passante utilisée)
- température processeurs, température du boîtier
- vitesse de rotation des ventillateurs
- sécurité (authentifications, tunnels)
- disponibilité des services (batch, interfaces)
- compteurs hardware si disponibles et pertinents (mcelog, . . .)
- sabordage de runs (limite de temps, plantages, bouclages)
- . . .

Spécificités

Que contrôler ?

Au niveau environnemental

- température de la salle (climatiseurs)
- état des onduleurs (% capacité, temps de disponibilité, état des batteries)
- temps d'attente en file par rapport au temps de calculs (cf. les logs du gestionnaire de batch, qacct)
- intrusions (réseaux, salle système)
- statistiques (taux d'utilisation du cluster, par utilisateur/groupe, facturations)
- niveau de disponibilité HA (ex. Road-Runner : 10mn d'arrêt par jour)
- ...

Spécificités

Que contrôler ?

Au niveau environnemental

- température de la salle (climatiseurs)
- état des onduleurs (% capacité, temps de disponibilité, état des batteries)
- temps d'attente en file par rapport au temps de calculs (cf. les logs du gestionnaire de batch, qacct)
- intrusions (réseaux, salle système)
- statistiques (taux d'utilisation du cluster, par utilisateur/groupe, facturations)
- niveau de disponibilité HA (ex. Road-Runner : 10mn d'arrêt par jour)
- ...

Spécificités

Que contrôler ?

Au niveau environnemental

- température de la salle (climatiseurs)
- état des onduleurs (% capacité, temps de disponibilité, état des batteries)
- temps d'attente en file par rapport au temps de calculs
  (cf. les logs du gestionnaire de batch, qacct)
- intrusions (réseaux, salle système)
- statistiques (taux d'utilisation du cluster, par utilisateur/groupe, facturations)
- niveau de disponibilité HA (ex. Road-Runner : 10mn d'arrêt par jour)
- ...

Spécificités

Que contrôler ?

Au niveau environnemental

- température de la salle (climatiseurs)
- état des onduleurs (% capacité, temps de disponibilité, état des batteries)
- temps d'attente en file par rapport au temps de calculs
  (cf. les logs du gestionnaire de batch, qacct)
- intrusions (réseaux, salle système)
- statistiques (taux d'utilisation du cluster, par utilisateur/groupe, facturations)
- niveau de disponibilité HA (ex. Road-Runner : 10mn d'arrêt par jour)
- ...

Spécificités

Que contrôler ?

Au niveau environnemental

- température de la salle (climatiseurs)
- état des onduleurs (% capacité, temps de disponibilité, état des batteries)
- temps d'attente en file par rapport au temps de calculs
  (cf. les logs du gestionnaire de batch, qacct)
- intrusions (réseaux, salle système)
- statistiques (taux d'utilisation du cluster, par utilisateur/groupe, facturations)
- niveau de disponibilité HA (ex. Road-Runner : 10mn d'arrêt par jour)
- ...

Spécificités

Que contrôler ?

Au niveau environnemental

- température de la salle (climatiseurs)
- état des onduleurs (% capacité, temps de disponibilité, état des batteries)
- temps d'attente en file par rapport au temps de calculs
  (cf. les logs du gestionnaire de batch, qacct)
- intrusions (réseaux, salle système)
- statistiques (taux d'utilisation du cluster, par utilisateur/groupe, facturations)
- niveau de disponibilité HA (ex. Road-Runner : 10mn d'arrêt par jour)
- ...

Spécificités

Que contrôler ?

Au niveau environnemental

- température de la salle (climatiseurs)
- état des onduleurs (% capacité, temps de disponibilité, état des batteries)
- temps d'attente en file par rapport au temps de calculs
  (cf. les logs du gestionnaire de batch, qacct)
- intrusions (réseaux, salle système)
- statistiques (taux d'utilisation du cluster, par utilisateur/groupe, facturations)
- niveau de disponibilité HA (ex. Road-Runner : 10mn d'arrêt par jour)
- ...

Monitoring de clusters

Plan

Monitoring de clusters

Outils et exemples

MRTG (Multi-Router Traffic Grapher)

http ://oss.oetiker.ch/mrtg/

- monitoring local, consolidation et affichage distants (RRD)
- basé sur SNMP et RRD tools
- scripts de configuration
- fabrique les graphes et les fichiers HTML

Monitoring de clusters

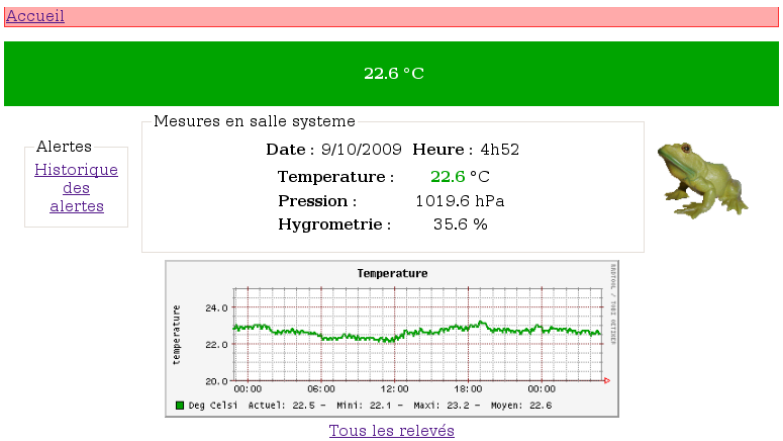
Outils et exemples

La Grenouille

http ://www.redge.net/frogd/fr/ : pilote Linux

http ://www.froggyhome.com/ : le capteur

- capteur (température-pression-hygrométrie) en salle système
- alimentée par USB/RS232
- pilote Linux (3 valeurs en fichier de log)
- adaptation avec RRD tools et des scripts



- automate d'extinction des machines lié (2 niveaux)

Monitoring de clusters

Outils et exemples

Surveillance des disques serveur

ftp ://ftp.sara.nl/pub/sara-omsa/dists/

- monitoring local invoqué à distance (CLI par ssh)
- interface de présentation + historisation
- carte PERC5i, PERC6i, PERC6E

Machine odin contrôleur controleur 1

Controller PERC 6/E Adapter (Slot 1)

Main System Chassis

	Fans	Memory	Power Supplies	Processors	Temperatures	Voltages	Batteries
0	Ok	Ok	Ok	Ok	Ok	Ok	Ok

Controllers

	ID	Status	Name	State
0	1	Ok	PERC 6/E Adapter	Ready

Connectors

	ID	Status	Name	State
0	0	Ok	Connector 0	Ready
1	1	Ok	Connector 1	Ready

Virtual Disks

	ID	Status	Name	State	Layout	Size	Device Name	Type
0	0	Ok	QCD1	Ready	RAID-5	1,862.00 GB (1999307276288 bytes)	/dev/sdb	SAS
1	1	Ok	COSMO1	Ready	RAID-5	1,862.00 GB (1999307276288 bytes)	/dev/sdc	SAS
2	2	Ok	QCD2	Ready	RAID-5	1,862.00 GB (1999307276288 bytes)	/dev/sdd	SAS
3	3	Ok	COSMO2	Ready	RAID-5	1,862.00 GB (1999307276288 bytes)	/dev/sde	SAS

Physical Disks

	ID	Status	Name	State	Failure Predicted	Type	Capacity	Hot Spare	Product ID	Used RAID Disk Space	Available RAID Disk Space
0	0:0:0	Ok	Physical Disk 0:0:0	Online	No	SAS	931.00 GB (999653638144 bytes)	No	ST3100064OSS	931.00 GB (999653638144 bytes)	0.00 GB (0 bytes)
1	0:0:1	Ok	Physical Disk 0:0:1	Online	No	SAS	931.00 GB (999653638144 bytes)	No	ST3100064OSS	931.00 GB (999653638144 bytes)	0.00 GB (0 bytes)
2	0:0:2	Ok	Physical Disk 0:0:2	Online	No	SAS	931.00 GB (999653638144 bytes)	No	ST3100064OSS	931.00 GB (999653638144 bytes)	0.00 GB (0 bytes)
3	0:0:3	Ok	Physical Disk 0:0:3	Online	No	SAS	931.00 GB (999653638144 bytes)	No	ST3100064OSS	931.00 GB (999653638144 bytes)	0.00 GB (0 bytes)
4	0:0:4	Ok	Physical Disk 0:0:4	Online	No	SAS	931.00 GB (999653638144 bytes)	No	ST3100064OSS	931.00 GB (999653638144 bytes)	0.00 GB (0 bytes)
5	0:0:5	Ok	Physical Disk 0:0:5	Online	No	SAS	931.00 GB (999653638144 bytes)	No	ST3100064OSS	931.00 GB (999653638144 bytes)	0.00 GB (0 bytes)
6	1:0:0	Ok	Physical Disk 1:0:0	Online	No	SAS	931.00 GB (999653638144 bytes)	No	ST3100064OSS	931.00 GB (999653638144 bytes)	0.00 GB (0 bytes)
7	1:0:1	Ok	Physical Disk 1:0:1	Online	No	SAS	931.00 GB (999653638144 bytes)	No	ST3100064OSS	931.00 GB (999653638144 bytes)	0.00 GB (0 bytes)

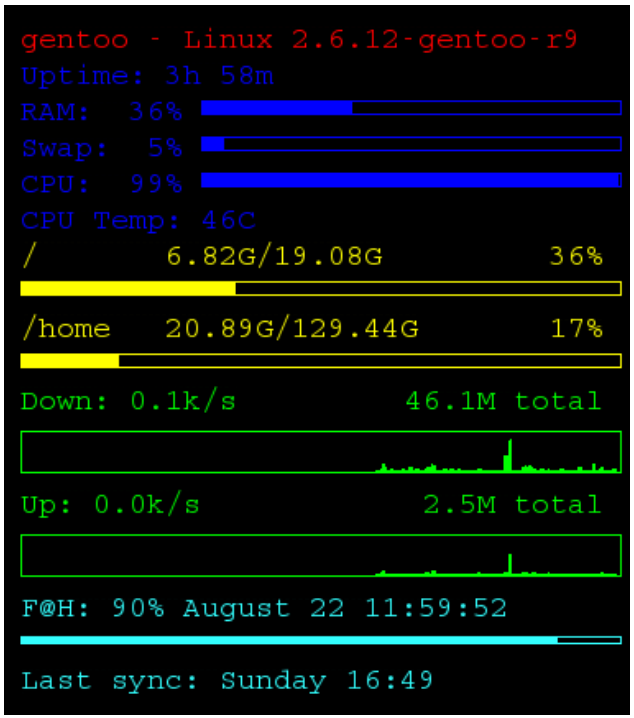
Monitoring de clusters

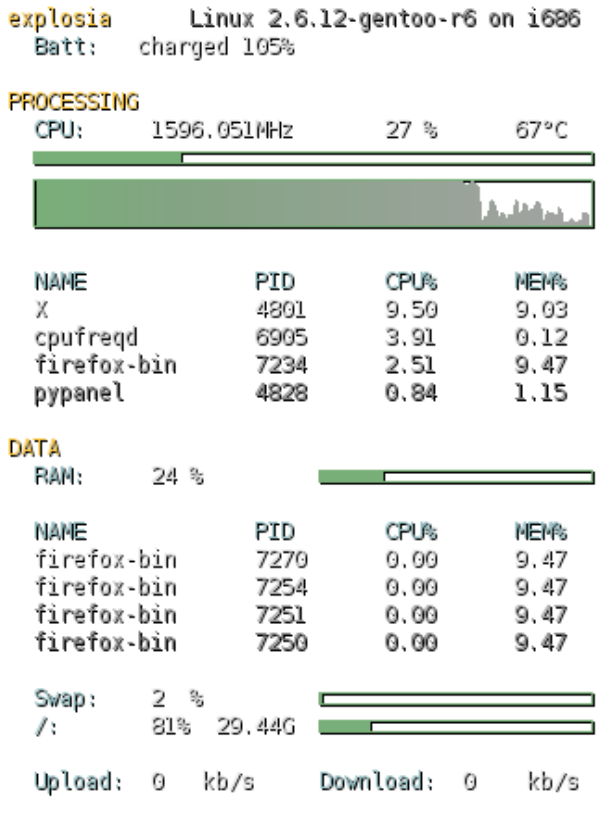
Outils et exemples

Conky : moniteur système graphique X (basé sur torsmo)

http ://conky.sourceforge.net/

- monitoring local affichage X distant
- largement configurable (250 objets internes, i2c, . . .)
- affichage graphique déporté (lourd en BP)





Monitoring de clusters

Outils et exemples

collectd : moteur de plugins écrit en C

- monitoring local performant (selon qualité des plug-ins)
- stockage local en RRD
- largement configurable (plugins, dont lm-sensors, . . .)
- de apache aux températures des HD

Monitoring de clusters

Outils et exemples

Ganglia : Monitoring hiérarchisé orienté cluster/grid

http ://ganglia.sourceforge.net/

- monitoring local (gmond) multi-plateformes
- très configurable (fréquences indépendantes par indicateur)
- communication entre les noeuds par multicast (disponibilité)
- consolidations intermédiaires (RRD) hiérarchiques (gmetad)
- affichage graphique PHP centralisé
- prévu pour monitorer les grappes de clusters jusqu'à 2000 nodes
- configuration *tricky* en fichier texte
  - gmond.conf pour les sondes
  - gmetad.conf pour les consolidations

Monitoring de clusters

Outils et exemples

Ganglia : Monitoring hiérarchisé orienté cluster/grid

http ://ganglia.sourceforge.net/

- monitoring local (gmond) multi-plateformes
- très configurable (fréquences indépendantes par indicateur)
- communication entre les noeuds par multicast (disponibilité)
- consolidations intermédiaires (RRD) hiérarchiques (gmetad)
- affichage graphique PHP centralisé
- prévu pour monitorer les grappes de clusters jusqu'à 2000 nodes
- configuration *tricky* en fichier texte
  - gmond.conf pour les sondes
  - gmetad.conf pour les consolidations

Monitoring de clusters

Outils et exemples

Ganglia : Monitoring hiérarchisé orienté cluster/grid

http ://ganglia.sourceforge.net/

- monitoring local (gmond) multi-plateformes
- très configurable (fréquences indépendantes par indicateur)
- communication entre les noeuds par multicast (disponibilité)
- consolidations intermédiaires (RRD) hiérarchiques (gmetad)
- affichage graphique PHP centralisé
- prévu pour monitorer les grappes de clusters jusqu'à 2000 nodes
- configuration *tricky* en fichier texte
  - gmond.conf pour les sondes
  - gmetad.conf pour les consolidations

Monitoring de clusters

Outils et exemples

Ganglia : Monitoring hiérarchisé orienté cluster/grid

http ://ganglia.sourceforge.net/

- monitoring local (gmond) multi-plateformes
- très configurable (fréquences indépendantes par indicateur)
- communication entre les noeuds par multicast (disponibilité)
- consolidations intermédiaires (RRD) hiérarchiques (gmetad)
- affichage graphique PHP centralisé
- prévu pour monitorer les grappes de clusters jusqu'à 2000 nodes
- configuration *tricky* en fichier texte
  - gmond.conf pour les sondes
  - gmetad.conf pour les consolidations

Monitoring de clusters

Outils et exemples

Ganglia : Monitoring hiérarchisé orienté cluster/grid

http ://ganglia.sourceforge.net/

- monitoring local (gmond) multi-plateformes
- très configurable (fréquences indépendantes par indicateur)
- communication entre les noeuds par multicast (disponibilité)
- consolidations intermédiaires (RRD) hiérarchiques (gmetad)
- affichage graphique PHP centralisé
- prévu pour monitorer les grappes de clusters jusqu'à 2000 nodes
- configuration *tricky* en fichier texte
  - gmond.conf pour les sondes
  - gmetad.conf pour les consolidations

Monitoring de clusters

Outils et exemples

Ganglia : Monitoring hiérarchisé orienté cluster/grid

http ://ganglia.sourceforge.net/

- monitoring local (gmond) multi-plateformes
- très configurable (fréquences indépendantes par indicateur)
- communication entre les noeuds par multicast (disponibilité)
- consolidations intermédiaires (RRD) hiérarchiques (gmetad)
- affichage graphique PHP centralisé
- prévu pour monitorer les grappes de clusters jusqu'à 2000 nodes
- configuration *tricky* en fichier texte
  - gmond.conf pour les sondes
  - gmetad.conf pour les consolidations

Monitoring de clusters

Outils et exemples

Ganglia : Monitoring hiérarchisé orienté cluster/grid

http ://ganglia.sourceforge.net/

- monitoring local (gmond) multi-plateformes
- très configurable (fréquences indépendantes par indicateur)
- communication entre les noeuds par multicast (disponibilité)
- consolidations intermédiaires (RRD) hiérarchiques (gmetad)
- affichage graphique PHP centralisé
- prévu pour monitorer les grappes de clusters jusqu'à 2000 nodes
- configuration *tricky* en fichier texte
  - gmond.conf pour les sondes
  - gmetad.conf pour les consolidations

Monitoring de clusters

Outils et exemples

Ganglia : Monitoring hiérarchisé orienté cluster/grid

http ://ganglia.sourceforge.net/

- monitoring local (gmond) multi-plateformes
- très configurable (fréquences indépendantes par indicateur)
- communication entre les noeuds par multicast (disponibilité)
- consolidations intermédiaires (RRD) hiérarchiques (gmetad)
- affichage graphique PHP centralisé
- prévu pour monitorer les grappes de clusters jusqu'à 2000 nodes
- configuration *tricky* en fichier texte
  - gmond.conf pour les sondes
  - gmetad.conf pour les consolidations

Monitoring de clusters

Outils et exemples

Ganglia : Monitoring hiérarchisé orienté cluster/grid

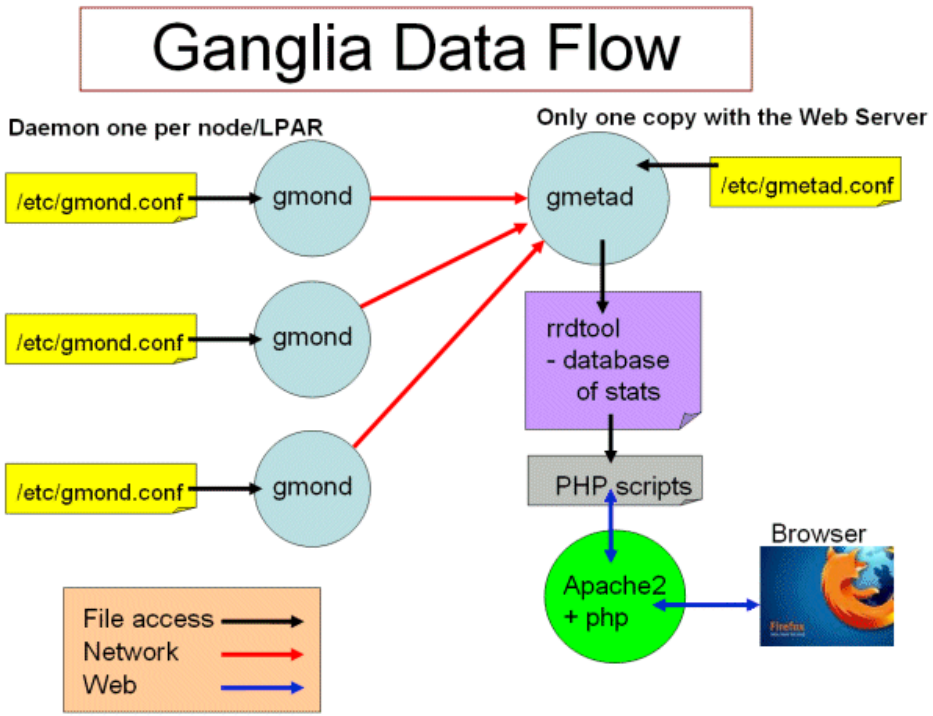
http ://ganglia.sourceforge.net/

- monitoring local (gmond) multi-plateformes
- très configurable (fréquences indépendantes par indicateur)
- communication entre les noeuds par multicast (disponibilité)
- consolidations intermédiaires (RRD) hiérarchiques (gmetad)
- affichage graphique PHP centralisé
- prévu pour monitorer les grappes de clusters jusqu'à 2000 nodes
- configuration *tricky* en fichier texte
  - gmond.conf pour les sondes
  - gmetad.conf pour les consolidations

Monitoring de clusters

Outils et exemples

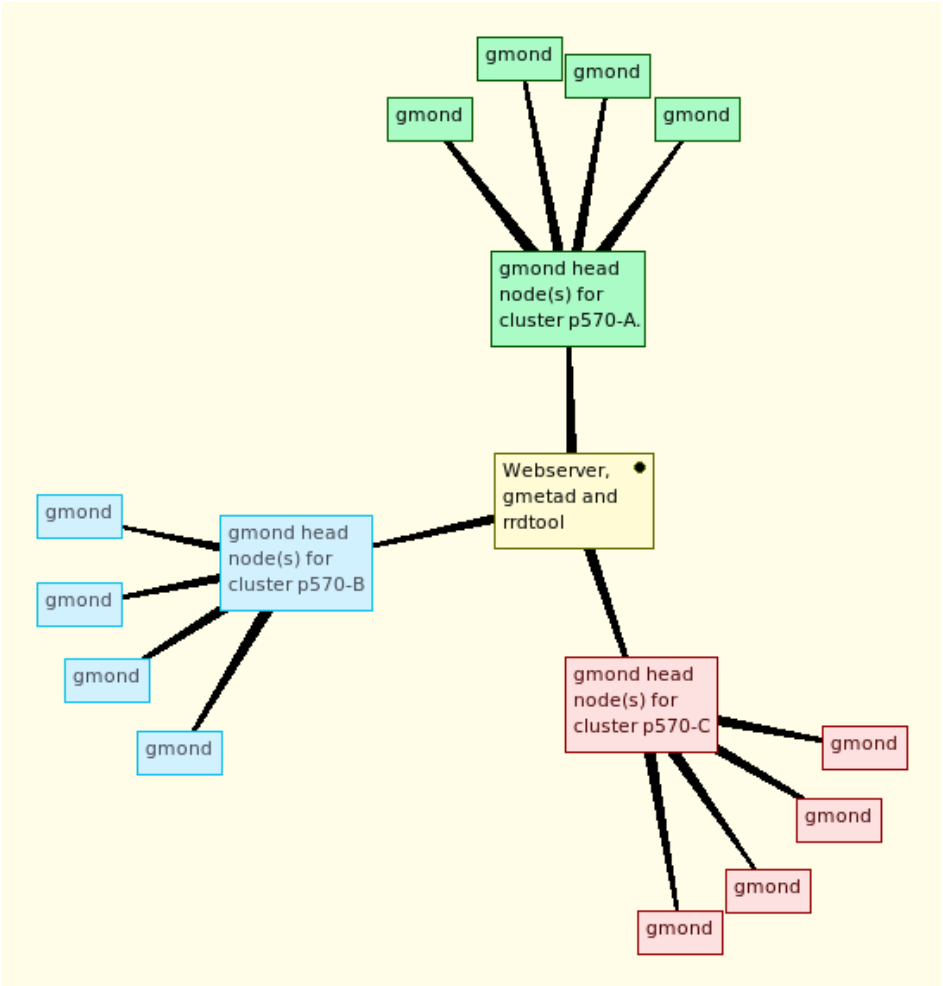
Ganglia : data flow



Monitoring de clusters

Outils et exemples

Ganglia : hiérarchie



Monitoring de clusters

Outils et exemples

Ganglia : exemples de paramètres

Server configuration

1. edit /etc/gmetad.conf, set gridname

```
gridname "p5 grid"
```

2. edit /etc/gmetad.conf, set sources:

```
# Add your head nodes to the data source for each cluster (machine)
data_source "p570-A" A-headnode1, A-headnode2
data_source "p570-B" B-headnode1, B-headnode2
# If p570-C is behind the firewall we have to add local ports
# which will be tunneled
data_source "p570-C" localhost:4662 localhost:4663
```

Client configuration

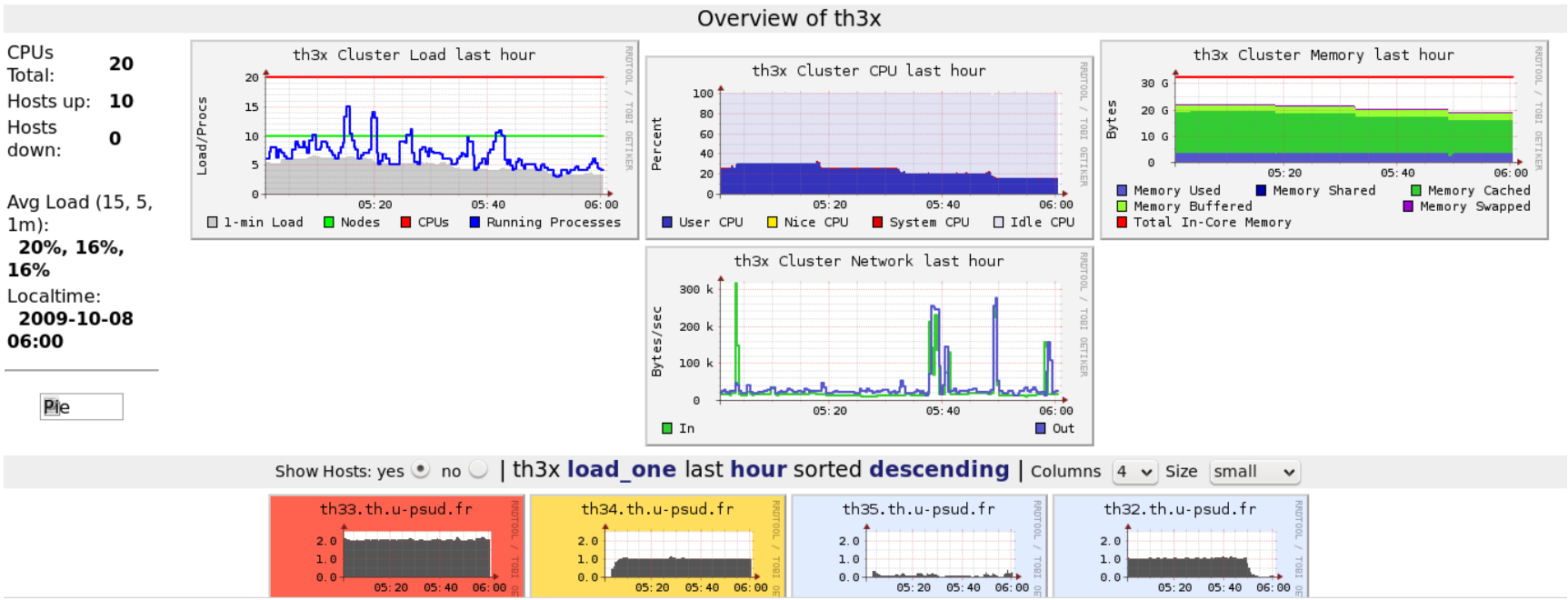
In each cluster (machine) cluster name and head node names should differ

```
globals {
  daemonize = yes
  setuid = yes
  user = root
  debug_level = 0
  max_udp_msg_len = 1472
  mute = no
  deaf = no
  host_dmax = 0 /*secs */
  cleanup_threshold = 300 /*secs */
  gexec = no
}
cluster {
  name = "p570-A"
  owner = ""
  latlong = "unspecified"
  url = "unspecified"
}
udp_send_channel {
 host = A-headnode1
 port = 8666
}
udp_send_channel {
 host = A-headnode2
 port = 8666
}
...
tcp_accept_channel {
  port = 8649
}
```

Monitoring de clusters

Outils et exemples

Ganglia : exemples de vues



Monitoring de clusters

Outils et exemples

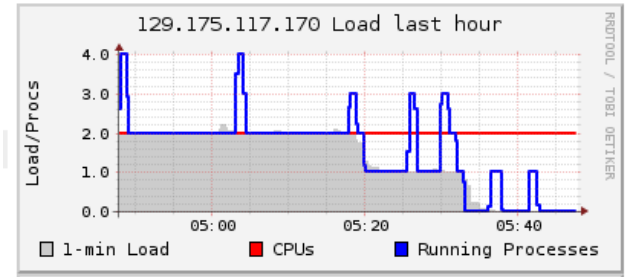
Ganglia : exemples de vues

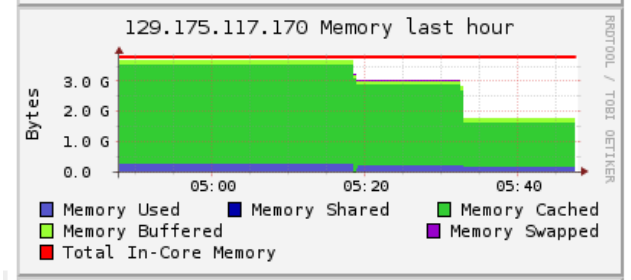
129.175.117.170 Overview

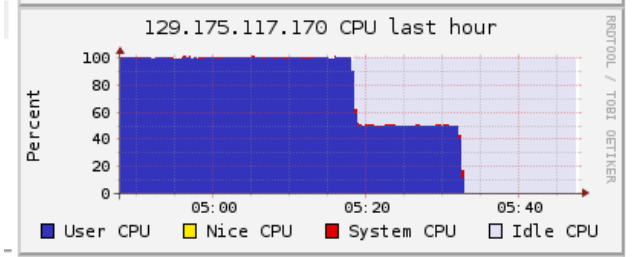
This host is up and running.

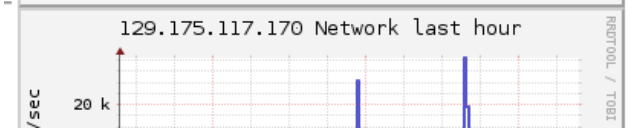
Time and String Metrics	
boottime	Wed, 02 Sep 2009 19:09:26 +0200
gexec	ON
Gmond Started	Wed, 07 Oct 2009 15:24:56 +0200
Last Reported	0 days, 0:00:06
machine_type	x86
os_name	Linux
os_release	2.6.27.29-0.1-pae
sys_clock	Wed, 07 Oct 2009 15:24:56 +0200
Uptime	35 days, 10:38:05

Constant Metrics	
cpu_num	2
cpu_speed	2600 MHz
mem_total	3965092 KB
mtu	1500 B
swap_total	1044216 KB





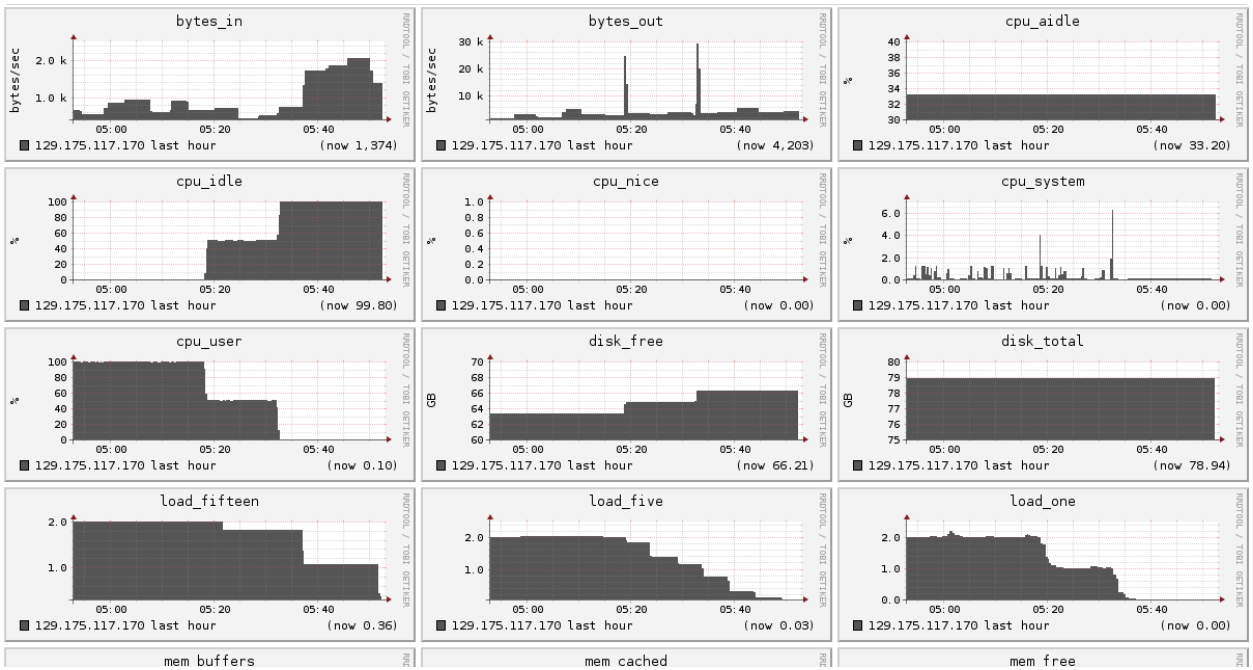




Monitoring de clusters

Outils et exemples

Ganglia : exemples de vues



Monitoring de clusters

Conclusion

le monitoring/supervision est nécessaire

beaucoup d'outils disponibles

choix adapté en fonction des configuration/contraintes

adopter une démarche qualité : évolutivité de la solution