**<u>Titre du livrable :</u> Rapport sur le support plate-forme avec indices de disponibilité (période 2008).**

**Partenaire émetteur : Bull**
**Auteur : Guy Hiot**
**Date : Décembre 2008**

# Table des matières

# 1 Le cluster Nova

L'objectif pour Bull en 2008 est de mettre à disposition des partenaires de POPS et des projets associés PARA et PARMA une configuration de cluster composé de serveurs (gros nœuds quadri-sockets) hébergeant des processeurs Intel Xeon quadri-cœurs.

Cette plate-forme **Nova** est destinée à permettre aux partenaires de Bull dans les différents projets de se familiariser avec les dernières versions de processeurs Intel Xeon, de porter et de paralléliser leurs applications sur cet environnement dans le but d'expérimenter et d'optimiser leurs algorithmes sur un grand nombre de cœurs afin d'anticiper les performances atteignables avec les futurs serveurs hautement parallèles développés par Bull dans le cadre de la convention POPS .

Cette plate-forme est située dans les locaux de Bull aux Clayes-sous-Bois, dans le département des Yvelines. Elle a aussi pour vocation de servir de vitrine technologique et, dans cette perspective, à être utilisée pour la réalisation de grands challenges nécessitant une puissance de calcul élevée.

Elle vient en complément de la plate-forme Fame2, constituée de nœuds hébergeant des processeurs Intel Itanium, investis précédemment dans le cadre du projet Fame2. La plate-forme Fame2 continue, pour le moment, d'être disponible pour les travaux des partenaires de POPS.
D'un point de vue administration système, les enseignements tirés de l'utilisation de Fame2 bénéficient à la mise en œuvre de Nova.

# 2 Eléments constitutifs du cluster

## *2.1 Nœuds du Cluster Nova*

Le cluster Nova (512 cœurs de calcul) est composé des équipements suivants:

- Un **serveur NovaScale R423** utilisé comme **nœud de Management (NV0)** du cluster. Ce nœud bi-sockets est équipé de processeurs Xeon E5430 quadri-cœurs à 2.66 GHz, de 16 GB de mémoire et d'un disque SAS de 146 GB. Le nom du cluster, reflété dans la base d'administration, est nv.

- Un **serveur NovaScale R423 (NV3)** utilisé comme **nœud d'entrée-sortie**. Ce nœud bi-sockets est équipé de processeurs Xeon E5430 quadri-cœurs à 2.66 GHz, de 32 GB de mémoire et de 2 disques SAS de 146 GB.

- Un **serveur NovaScale R423 (NV2)** utilisé comme **nœud MDS**. Ce nœud bi-sockets est équipé de processeurs Xeon E5430 quadri-cœurs à 2.66 GHz, de 32 GB de mémoire, de 2 disques SAS de 146 GB et 2 disques SAS de 300 GB.

- **Seize nœuds de calcul (NV4 à NV19) NovaScale R480-E1 quadri-sockets**, chaque socket étant équipée d'un processeur Xeon E7350 Tigerton quadri-cœur à 2.93 GHz et 2x4 MB de cache L2. Chaque nœud dispose de 48 GB de mémoire et de 1 disque SAS de 146 GB 10 Krpm.

- **Seize nœuds de calcul (NV20 à NV35) NovaScale R480-E1 quadri-sockets**, chaque socket étant équipée d'un processeur Xeon E7310 Tigerton quadri-cœur à 1.6 GHz et 2x2 MB de cache L2. Chaque nœud dispose de 48 GB de mémoire, d'un FSB à 1066 MHz et d'un disque SAS de 146 GB 10 Krpm.

## *2.2 Interconnect*

Tous ces noeuds sont interconnectés en **Infiniband** (20 GB/s) à l'aide d'une fabrique composée de switches Voltaire type ISR 9024D. Les cartes IB sont des Mellanox ConnectX.

Les nœuds sont également tous reliés à un réseau **Giga-bit Ethernet** à l'aide d'un switch Cisco 2960G équipé de 44 ports.

## *2.3 Stockage*

Outre le stockage local aux nœuds, deux baies disque FDA 2500, équipées de 210 disques de 300 GB organisés en RAID 5, offrent une capacité utile de stockage externe de **49,2 Téraoctets**.

Ces baies sont interfacées en **Fiber Channel** avec le nœud d'entrée-sortie. Chacune de ces baies possède deux Storage Processors équipés de 8 GB de mémoire cache et 4 ports FC à 4 Gbits.

Le nœud d'E/S possède 8 cœurs ce qui lui permet de gérer efficacement les nombreux petits paquets IP qui sont générés quand **NFS** est utilisé.
Quand ce nœud est configuré avec **Lustre** (système de gestion de fichiers parallélisé), un autre nœud est utilisé pour MDS (Metadata server).

L'espace disque en stockage externe devrait être doublé au 4ème trimestre 2008.

## *2.4 Installation et mise à disposition de Nova*

Le cluster Nova a été investi par Bull au 1er semestre 2008. Les différents éléments matériels ont été livrés dans la période avril-mai.

Plusieurs difficultés ont été rencontrées lors des tests d'installation et notamment un problème de mauvaise configuration du BIOS des R480 qui perturbait le réseau Ethernet d'administration et empêchait de transférer les images sur les nœuds de calcul.
Fin juin, le cluster fonctionnait avec le système d'exploitation XBAS5 V1.1. La version de MPIBull2 installée est la 1.3.7.1.

Ensuite, il a fallu installer, configurer, mettre au point les scripts et tester, dans différents cas d'utilisation (et notamment avec des travaux utilisant OPENMP ou MPI ), l'environnement d'exploitation **Pbspro** qui sera obligatoirement utilisé par les partenaires pour s'allouer des ressources, soumettre leurs travaux (commande qsub) et contrôler leur exécution.

Les connexions externes à Nova, depuis Internet, s'effectuent en utilisant SSH. Les partenaires, qui avaient accès à Fame2, ont automatiquement accès à Nova et retrouvent leur environnement de travail (home_nfs) qui est implanté sur une ressource disque partagée entre les deux clusters.

Finalement, **l'ouverture de l'accès au cluster Nova a été officiellement annoncée aux partenaires le 31 juillet 2008**.

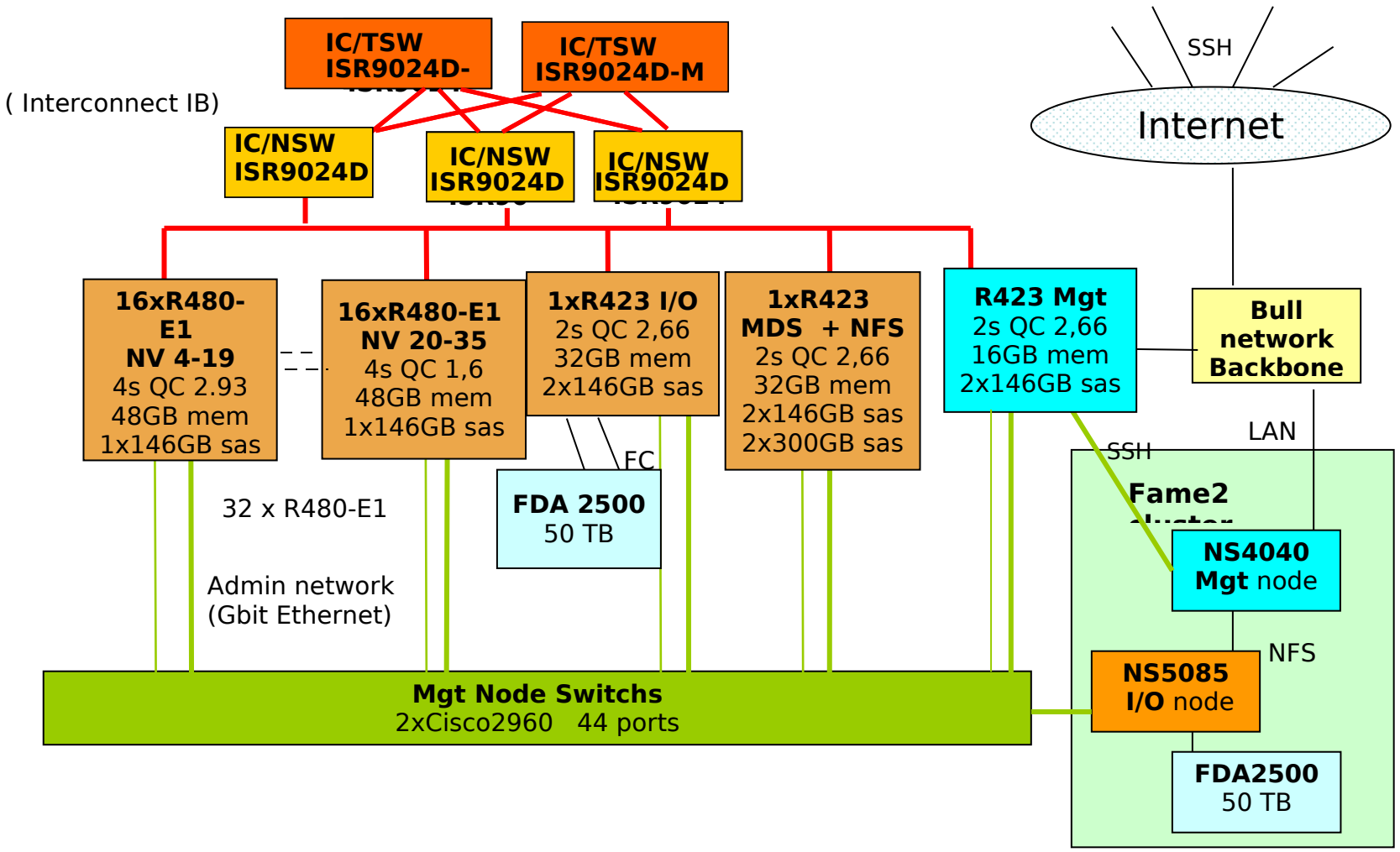

IC/TSW ISR9024D-
IC/TSW ISR9024D-M
( Interconnect IB)
IC/NSW ISR9024D
IC/NSW ISR9024D
IC/NSW ISR9024D
SSH
Internet
16xR480-E1 NV 4-19 4s QC 2.93 48GB mem 1x146GB sas
16xR480-E1 NV 20-35 4s QC 1,6 48GB mem 1x146GB sas
1xR423 I/O 2s QC 2,66 32GB mem 2x146GB sas
1xR423 MDS + NFS 2s QC 2,66 32GB mem 2x146GB sas 2x300GB sas
R423 Mgt 2s QC 2,66 16GB mem 2x146GB sas
Bull network Backbone
LAN
SSH
FC
FDA 2500 50 TB
32 x R480-E1
Admin network (Gbit Ethernet)
Fame2
NS4040 Mgt node
NFS
NS5085 I/O node
FDA2500 50 TB
Mgt Node Switchs 2xCisco2960 44 ports

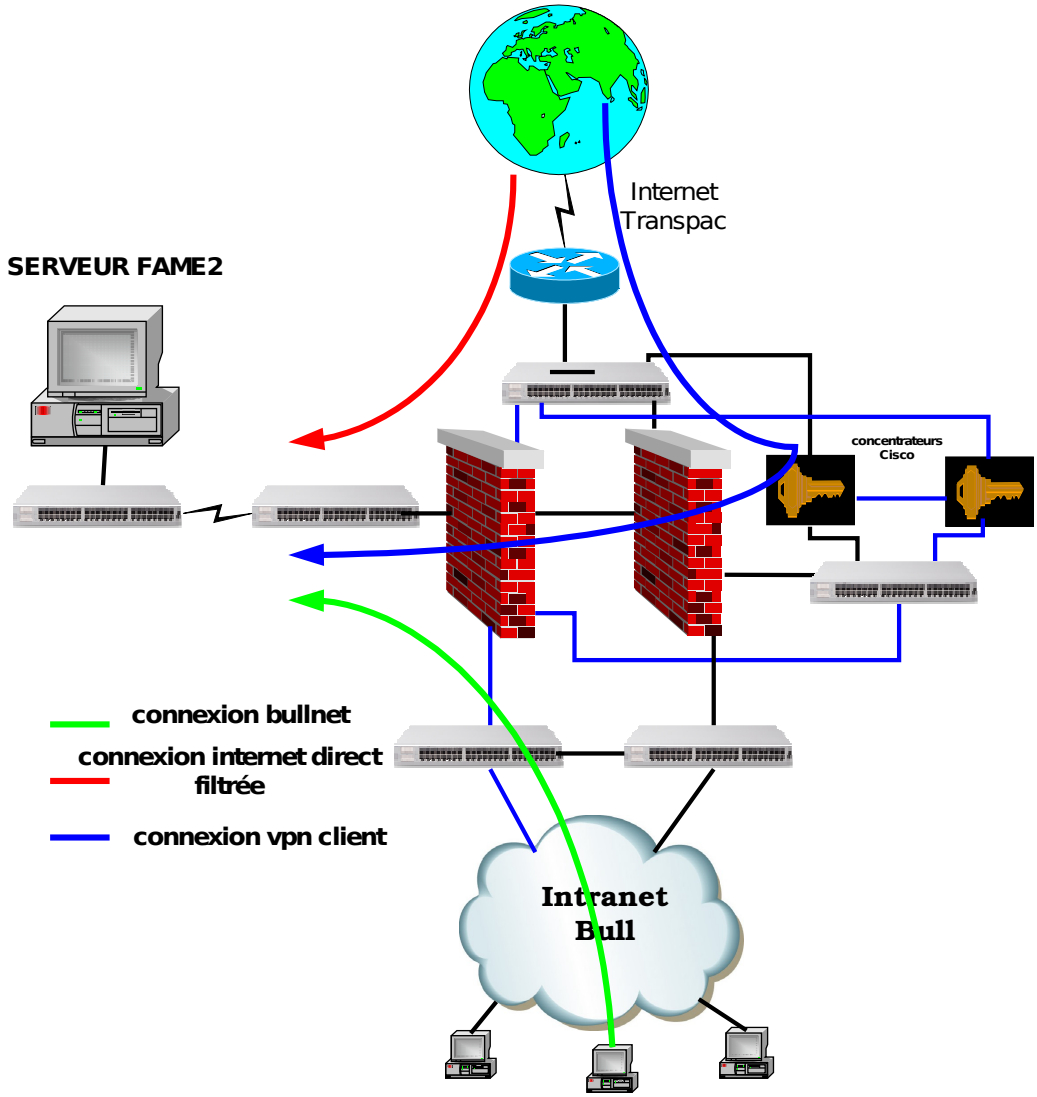
Figure 1-1

# 3 Accès réseau

Comme pour Fame2, tous les partenaires accèdent le cluster Nova à travers Internet.

## *3.1 Structure réseau d'accueil des partenaires*

Rappel de l'architecture réseau mise en place:




Le choix préconisé au départ était d'utiliser autant que possible des connexions VPN sécurisées.

La mise en place de connexion VPN posant problème chez certains partenaires à cause de leur architecture réseau et leurs règles de sécurité, on autorise le plus souvent des connexions SSH directes.

Ces connexions, by-passant les serveurs d'accès VPN, ont besoin d'être identifiées nominativement au niveau du pare-feu d'accès de Bull. Les utilisateurs désireux de travailler dans ce mode doivent alors décliner leur adresse IP source. Un délai de plusieurs jours est alors nécessaire pour que France Télécom ajoute sélectivement une nouvelle règle d'accès au niveau du pare-feu.

## *3.2 Gestion des demandes de Connexions*

Toutes les demandes de connexion sont faites en utilisant un formulaire qui a été réactualisé en 2008 à l'occasion de Nova. Une version anglaise a été mise à la disposition des partenaires européens du projet Parma.

# 4 Utilisation du cluster Nova

Les administrateurs suivent l'utilisation du cluster à l'aide de l'outil de soumission de travaux et d'instrumentation PBS Pro.
Voici des exemples de synthèses correspondant aux mois d'octobre et de novembre 2008.

PBS Pro Cluster Accounting Summary Statistics

Report from **Wed Oct 1 2008 00:00:00 to Fri Oct 31 2008 23:59:59**

| | | | En secondes | | | | En heures | |
|---|---|---|---|---|---|---|---|---|
| Group | Username | # of jobs | Total CPU Time | Total Wall Time | Efcy. | Average Wait Time | Total CPU Time | Total Wall Time |
| bull | bastec | 8 | 0 | 4 353 | 0.000 | 2 | 0 | 1 |
| bull | bourgeod | 1 | 0 | 23 | 0.000 | 2 | 0 | 0 |
| bull | calegarp | 1 | 6 472 | 883 | 7.330 | 2 | 2 | 0 |
| bull | moreljm | 12 | 2 | 40 011 | 0.000 | 2 | 0 | 11 |
| bullcomet | gueloujl | 1 | 246 556 | 82 197 | 3.000 | 0 | 68 | 23 |
| bullcomet | kobenaj | 73 | 224 | 218 | 1.028 | 1 | 0 | 0 |
| bullcomet | senglongt | 5 | 678 890 | 255 427 | 2.658 | 310071 | 189 | 71 |
| bullcomet | valleef | 3 | 236 120 | 78 758 | 2.998 | 0 | 66 | 22 |
| bullechir | berthelc | 11 | 204 953 | 22 908 | 8.947 | 0 | 57 | 6 |
| dassault | dinhq | 16 | 3 433 | 1 893 | 1.814 | 0 | 1 | 1 |
| ecp | venetc | 111 | 1 291 648 | 1 355 832 | 0.953 | 0 | 359 | 377 |
| fzj | hermannm | 1 | 0 | 1 716 | 0.000 | 2 | 0 | 0 |
| fzj | mohrb | 27 | 1 725 | 44 302 | 0.039 | 2 | 0 | 12 |
| gns | menzelr | 89 | 3 947 586 | 776 568 | 5.083 | 80 | 1 097 | 216 |
| gwt | machc | 25 | 1 975 | 1 972 653 | 0.001 | 1 | 1 | 548 |
| gwt | williamt | 23 | 27 | 39 237 | 0.001 | 1 | 0 | 11 |
| hlrs | dichevk | 69 | 898 422 | 522 249 | 1.720 | 18 | 250 | 145 |
| hlrs | himmlerv | 42 | 10 830 | 151 894 | 0.071 | 143 | 3 | 42 |
| it-sudparis | mullera | 3 | 309 | 3 143 | 0.098 | 1 | 0 | 1 |
| it-sudparis | parrotc | 1 458 | 530 | 63 373 | 0.008 | 0 | 0 | 18 |
| medit | doppelto | 25 | 363 547 | 120 869 | 3.008 | 2 | 101 | 34 |
| opsim | allainjc | 8 | 39 427 | 2 812 | 14.021 | 1 | 11 | 1 |
| opsim | chatill2 | 83 | 31 399 945 | 2 373 521 | 13.229 | 6124 | 8 722 | 659 |
| opsim | maillot2 | 1 | 0 | 4 146 | 0.000 | 2 | 0 | 1 |
| pbs | pbs | 1 | 0 | 188 | 0.000 | 0 | 0 | 0 |
| ptu | zuckerms | 12 | 258 430 | 2 974 095 | 0.087 | 72 | 72 | 826 |
| recom | risiob | 196 | 20 363 362 | 1 109 704 | 18.350 | 586 | 5 656 | 308 |
| renault | chatillm | 10 | 616 675 | 79 445 | 7.762 | 344 | 171 | 22 |
| renault | maillots | 50 | 100 191 | 80 363 | 1.247 | 187 | 28 | 22 |
| renault | masc | 3 | 204 961 | 14 655 | 13.986 | 2 | 57 | 4 |
| renault | sidorkie | 390 | 67 489 662 | 8 635 823 | 7.815 | 79458 | 18 747 | 2 399 |
| resonate-mp4 | souchayp | 3 | 408 | 670 | 0.609 | 1 | 0 | 0 |
| uvsq | jaegerj | 1 | 15 | 1 262 | 0.012 | 2 | 0 | 0 |
| zih | mixh | 3 | 5 | 34 750 | 0.000 | 2 | 0 | 10 |
| cea | perrottl | 1 | 0 | 197 | 0.000 | 2 | 0 | 0 |
| | TOTAL | 2 766 | 128 366 330 | 20 850 138 | 6.157 | 12000 | 35 657 | 5 792 |

PBS Pro Cluster Accounting Summary Statistics

**Report from Sat Nov 1 2008 00:00:00 to Sun Nov 30 2008 23:59:59**

**Note: All times displayed in seconds.**

| Group | Username | # of jobs | Total CPU Time | Total Wall Time | Efcy. | Average Wait Time |
|---|---|---|---|---|---|---|
| bull | bastec | 2 | 0 | 9 | 0.000 | 2 |
| bull | bienatip | 7 | 1 403 | 114 329 | 0.012 | 1731 |
| bull | bourgeod | 7 | 0 | 3 737 | 0.000 | 2 |
| bull | jeaugeys | 13 | 2 993 | 1 899 | 1.576 | 1 |
| bull | moreljm | 1 | 0 | 452 | 0.000 | 3 |
| bullcomet | allojp | 2 | 0 | 53 | 0.000 | 2 |
| bullcomet | gueloujl | 4 | 0 | 260 957 | 0.000 | 2 |
| bullcomet | kobenaj | 86 | 441 | 731 | 0.603 | 1 |
| bullcomet | senglongt | 1 | 0 | 899 | 0.000 | 3 |
| dassault | carayolq | 3 | 0 | 136 | 0.000 | 2 |
| dassault | gounote | 13 | 4 | 4 916 | 0.001 | 1 |
| ecp | venetc | 63 | 213 526 | 59 981 | 3.560 | 0 |
| fzj | mohrb | 5 | 309 | 19 137 | 0.016 | 2 |
| gns | menzelr | 93 | 2 548 227 | 453 260 | 5.622 | 1 |
| gwt | williamt | 31 | 124 | 105 115 | 0.001 | 1 |
| hlrs | himmlerv | 22 | 2 229 732 | 522 278 | 4.269 | 1 |
| it-sudparis | mullera | 3 | 7 | 495 | 0.014 | 2 |
| it-sudparis | parrotc | 81 | 34 | 73 | 0.466 | 0 |
| magma | lukaszek | 22 | 507 929 | 244 300 | 2.079 | 0 |
| medit | doppelto | 107 | 68 349 753 | 4 954 713 | 13.795 | 26 |
| opsim | allainjc | 6 | 1 491 859 | 134 336 | 11.105 | 2 |
| opsim | chatill2 | 79 | 52 184 069 | 4 353 903 | 11.986 | 24106 |
| ptu | koliais | 8 | 5 328 | 1 371 274 | 0.004 | 2 |
| ptu | zuckerms | 13 | 226 | 2 078 196 | 0.000 | 2 |
| recom | risiob | 2 | 2 464 994 | 352 124 | 7.000 | 2 |
| renault | chatillm | 17 | 7 480 | 1 421 | 5.264 | 10 |
| renault | masc | 15 | 211 028 | 31 978 | 6.599 | 343 |
| renault | sidorkie | 35 | 4 894 032 | 619 749 | 7.897 | 164916 |
| zih | mixh | 2 | 0 | 25 507 | 0.000 | 2 |
| | TOTAL | 743 | 135 113 498 | 15 715 958 | 8.597 | 10359 |

# 5 Surveillance et Disponibilité du cluster Nova

Voici le tableau de bord des interventions et incidents relatifs à Nova en 2008 :

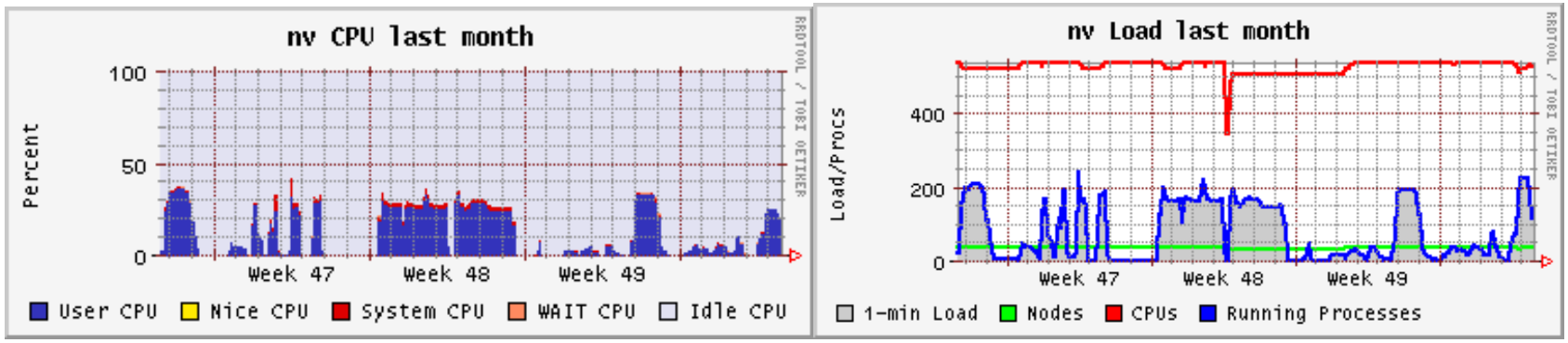
| Date | Evènements | Actions | Problèmes rencontrés | Pb résolu le |
|---|---|---|---|---|
| | | JUIN | | |
| 02/06/2008 | Arrivée du cluster d'Angers | Montage du cluster Nova | | |
| | | | | |
| | Impossible de transférer une image sur les nœuds de calcul | Reconfiguration du BIOS avec les infos données par ECH | la configuration du BIOS n'est pas bonne | |
| | | | | |
| 20/06/2008 | NSMaster remonte des alertes sur NV26 | Changement de la carte SIMSO et configuration | la cmd IPMITOOLS ne fonctionne pas sur nv26, carte SIMSO HS | 03/07/2008 |
| 25/06/2008 | voyant allumé en face avant de nv15 | Changement du bloque alim | Un bloque d'alimentation est HS | 03/07/2008 |
| 26/06/2008 | NV30 est rouge dans ganglia | Changement de la carte mère HS + configuration du BIOS | | 03/07/2008 |
| | | JUILLET | | |
| | | | | |
| 02/07/2008 | installation de pbspro | Installation d'un serveur PBSPRO et configuration de l'interconnexion du réseau entre PBSPRO et le cluser | le réseau n'est pas configuré correctement | 04/07/2008 |
| 15/07/2008 | | reboot de nv30 avec maj de la mémoire dans le BIOS | PB mémoire sur nv23 et 30 | |
| 16/07/2008 | NV23 est rouge dans ganglia | Changement de la carte mère HS | Impossible de rebouter NV23 | |
| | | | | |

| | | | | |
|---|---|---|---|---|
| | | AOUT | | |
| | ouverture du cluster aux partenaires | | | |
| | demande du CEA pour accèder à 1 TB | configuration de fda2500 | FDA2500 n'est pas configuré | |
| | | | | |
| 20/08/2008 | M Kirill signal que NV17 est lent | la cmd ipmitool remonte des erreur bus, demande au support (M piccoli d'annalyser) | Lors d'un calcul NV17 est plus lent que les autres nœuds | |
| | | | | |
| 28/08/2008 | impossible de se connecter à IPMITOOL sur le nœud NV30 | Changement du mot de passe de IPMITOOL | pas de log hardware de nv30 | |
| | | SEPTEMBRE | | |
| | | | | |
| 04/09/2008 | Ganglia signal un manque de mémoire dans nv17 | reboot et reset des mémoires dans le bios | | |
| 10/09/2008 | ajout d'un nouveau système | reconfiguration des ports | impossible de connecter un nouveau système car les ports disponibles du switch sont lockés | |
| | | | | |
| 29/09/2008 | Ganglia signale qu'il manque un cpu dans nv17 | reboot et reset des cpu dans le bios | | |
| | | | | |
| | | OCTOBRE | | |
| | | | | |
| 06/10/2008 | Ganglia signal un manque de mémoire dans nv23 | reboot et reset des mémoires dans le bios | | |
| | | NOVEMBRE | | |
| | | | | |
| | | | | |
| | | DECEMBRE | | |
| | | | | |

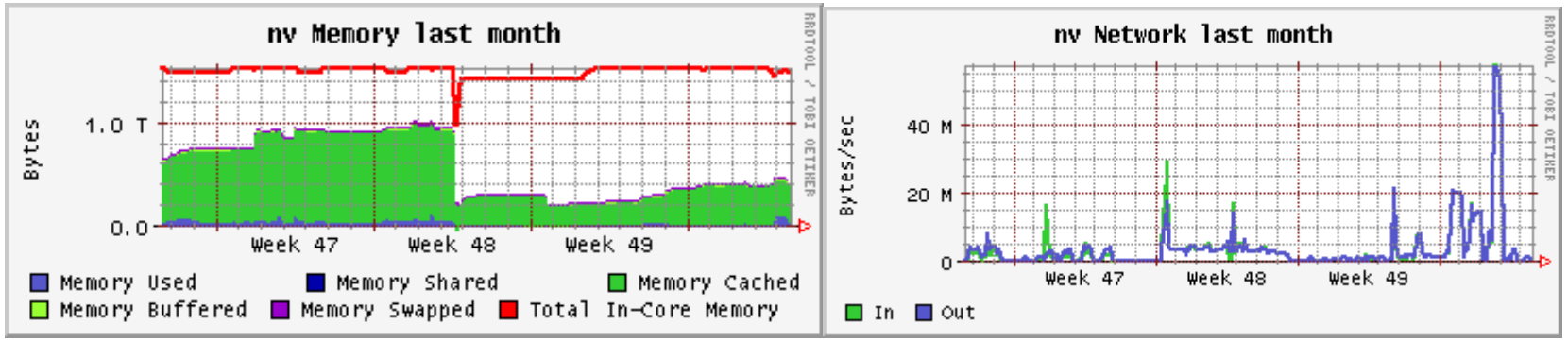
| | | | | |
|---|---|---|---|---|
| 09/12/2008 | Ganglia signal un manque de mémoire dans nv11 | reboot et reset des mémoires dans le bios | | |

La surveillance et la disponibilité globale du cluster ainsi que de chacun des nœuds sont suivis avec l'outil Ganglia.

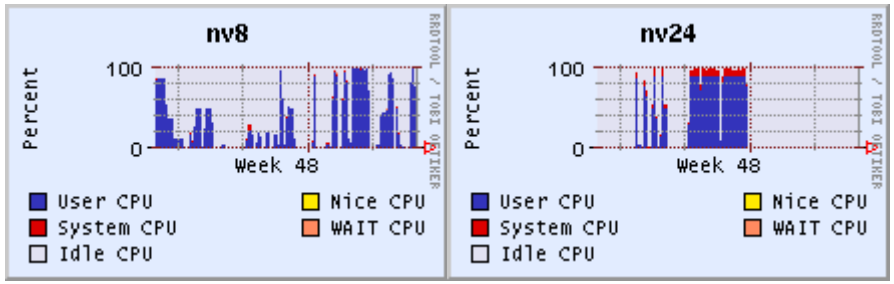
Voici quelques exemples illustrant la surveillance globale :

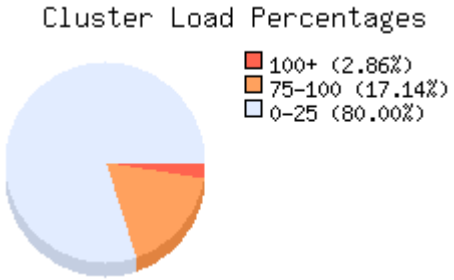



Surveillance globale

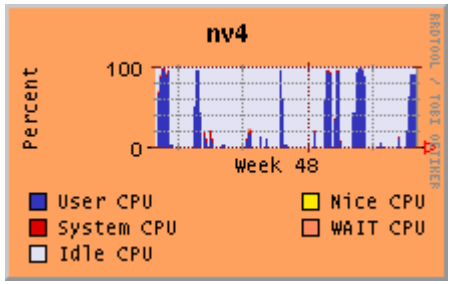


et des exemples visualisant l'activité et la disponibilité individuelle des nœuds :

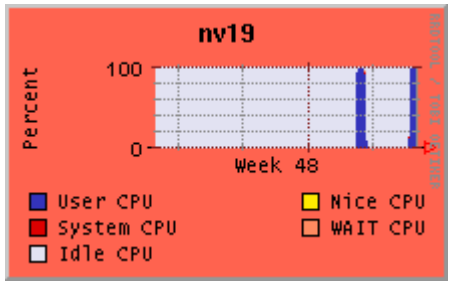



Le fond bleu clair indique que ces nœuds sont actuellement utilisés à moins de 25%.
Les graphiques à l'intérieur visualisent l'activité pour le mois.




Le fond orange indique que le nœud nv4 est utilisé entre 75 et 100%




Le fond rouge indique que le nœud nv19 est utilisé entre à 100%

Les bandes blanches que présentent les nœuds nv25 et nv30 correspondent à des périodes d'indisponibilité de ces nœuds.