Panache Conseil

Adaptabilité des infrastructures de communication face aux déluges (de données)

Dany.Vandromme@panache-conseil.fr

Panache Conseil

Résumé

- Les réseaux de télécommunications pour la recherche et l'éducation (REN) ont toujours eu à faire face à des besoins extrêmes exprimés par les utilisateurs. Ces défis permanents sont relevés en tirant le meilleur profit de la technologie, qui est caractérisée aujourd'hui par le caractère hybride des infrastructures. Cela permet en particulier de distinguer la topologie des infrastructures physiques vis-à-vis des architectures logiques mises en place. Cela permet un certain niveau d'adaptabilité dont la plus récente illustration est la mise en place du réseau LHCONE sur un temps beaucoup plus court que celui qui a mené au déploiement du réseau LHCOPN. A la suite de cette illustration, d'autres exemples seront donnés pour illustrer la prise en compte du déluge annoncé, en radioastronomie et en biologie par exemple, sans occulter néanmoins d'autres préoccupations liées à l'économie et à l'environnement.

Panache Conseil

Du LHCOPN au LHCONE

- LHC: Large Hadron Collider
- Nouvel accélérateur de particules installé dans le tunnel circulaire du CERN (26 km diamètre) après le démontage du LEP
- 4 Expériences sont installées sur le LHC: ATLAS, ALICE, CMS et LHCb.
- Chaque expérience dispose de ses propres détecteurs et chaînes de traitement de données.

Panache Conseil

Du LHCOPN au LHCONE

- Les quatre expériences partagent la même « machine » pour fabriquer les collisions.
- Dès le début de la construction du LHC, la communauté HEP prévoyait la production de volumes de données beaucoup plus importants que tout ce qui avait jamais été réalisé précédemment (> 15 Po/an) !
- Le Conseil du CERN (assemblée des Membres de l'organisation) a décidé qu'il ne revenait pas au CERN de supporter la gestion de l'ensemble des données produites!

Panache Conseil

Du LHCOPN au LHCONE

- Parce que:
  - Les utilisateurs des données produites ne sont pas au CERN (~10 000 physiciens répartis dans le monde entier)
  - Le coût de gestion des données dépasse les moyens financiers et techniques du laboratoire CERN.
- Il a donc été décidé de « placer » les données au plus proche des utilisateurs!
- Ce placement vaut pour le transport (LHCOPN), le stockage (T0/T1/T2/T3) et pour le traitement numérique (W-LCG).

Panache Conseil

Du LHCOPN au LHCONE

- Modèle de placement déterministe des données → MONARC
- Les données issues des détecteurs sont traitées localement avant d'être stockées dans le Tier-0 (CERN)
- Le Tier-0 sert les Tier-1 qui sont répartis dans 11 centres (7 en Europe, 2 aux US, 1 au Canada et 1 à Taiwan).
- Les Tier-1 servent de relais de stockage, hébergeant tout ou partie des données d'une ou plusieurs expériences.

Panache Conseil

Du LHCOPN au LHCONE

- Les Tier-1 servent les Tier-2 (≥ 75 dans le monde). Les Tier-2 sont utilisés pour le stockage des données, et éventuellement l'hébergement des ressources de calcul partagées.
- Les Tier-2 servent eux-mêmes les Tier-3 où sont les utilisateurs des données. Les Tier-3 stockent les données et les moyens de traitement.

Panache Conseil

Du LHCOPN au LHCONE

- Dans le modèle MONARC, les données sont pré-placées dans les Tier-2 et les Tier-3!
- Chaque Tier-1 sert ses propres Tier-2. Les Tier-2 n'utilisent que les données de leur Tier-1 de référence.
- Tous les Tier-1 sont raccordés au Tier-0! Les Tier-1 sont interconnectés entre eux (sans passer nécessairement par le CERN…).

Panache Conseil

Du LHCOPN au LHCONE

- Dès la fin des années 90s, HEP a mis au défi les NREN de satisfaire les besoins annoncés…
- Le débat a été long, parfois tendu, toujours formateur, mais au final la technologie a progressé suffisamment vite pour que les NREN soient au rendez-vous des périodes d'essai et de validation ("challenges") puis de la mise en route du LHC!

Panache Conseil

Du LHCOPN au LHCONE

- LHCOPN (LHC Optical Private Network) a été déployé via un ensemble de longueurs d'ondes T0/T1 (10 Gb/s) mises à la disposition des Tier-1 par les NREN et GEANT (ainsi que ESNET).
- En complément, les NREN ont fourni les circuits de backup/contournement T1/T1 (par exemple CC-IN2P3-Lyon ←→ GridKA-Karlsruhe par CBF via Strasbourg).

Panache Conseil

Du LHCOPN au LHCONE

- Dans le modèle initial, LHCOPN ne concerne que Tier-0 et Tier-1.
- Tier-1/Tier-2 (et à fortiori Tier-3) ne relève que des NREN dans leur périmètre national via un service IP suffisamment provisionné!
- Rappel pour la France: Tier-1: CC-IN2P3 à Lyon, Tier-2: GRIF (idF), Nantes, Clermont, Strasbourg, Grenoble, Marseille, Annecy et Lyon.
- Tier-1-FR a aussi des liens privilégiés avec les Tier-2 de Beijing et Tokyo!

Panache Conseil

Du LHCOPN au LHCONE

- Après le démarrage du LHC, le constat est clair: Les Tier-2 échangent beaucoup de données entre eux, et avec beaucoup de Tier-1.
- Le modèle souhaité n'est plus MONARC! Tous les Tier-2 échangent avec tous les Tier-1 et tous les Tier-2!
- La réponse "Full Mesh" n'est pas envisageable pour ~100 nœuds!

Panache Conseil

Du LHCOPN au LHCONE

- Les infrastructures de transit IP permettent d'acheminer tous les trafics, mais au détriement du trafic des autres communautés scientifiques.
- Par exemple, le LLR occupe jusqu'à 9 Gb/s sur une interface 10 Gb/s RENATER/GEANT pour dialoguer avec un seul laboratoire Tier-2 de Los Angeles…

Panache Conseil

Du LHCOPN au LHCONE

- La réponse, élaborée conjointement par les NREN et la communauté HEP, consiste à déployer un VPN-L3 au niveau global (NREN + GEANT + ESNET + Internet2) dédié aux trafics T2/T2 et T2/T1.
- En Europe, les NREN raccordent à la fois les Tier-1 et les Tier-2: “facile”!
- Aux US, les Tier-1 sont sur Esnet, et les Tier-2 sont sur Internet2: Pas si “facile”!

Panache Conseil

Du LHCOPN au LHCONE

- Les physiciens exigent un service Premium! Double connectivité des Tier-2 sur leur NREN, Double connectivité de LHCONE-FR sur LHCONE-GEANT, Double connectivité de LHCONE-GEANT sur LHCONE-US, 10Gb/s minimum par circuit…
- La réponse européenne est portée initialement par GEANT, DFN (Allemagne), RedIRIS (Espagne), GARR (Italie) et RENATER (France), mais le LHCONE est ouvert à tous les NREN raccordant au moins 1 Tier-2 (ou Tier-1).

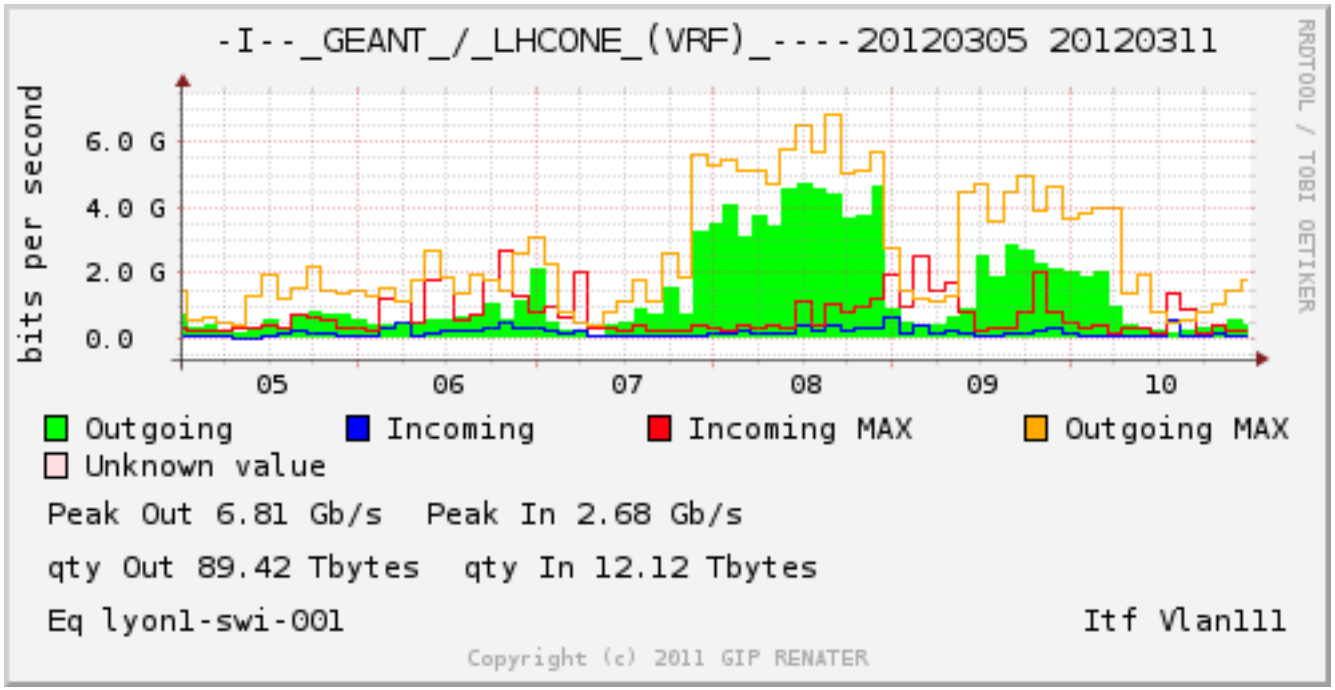
Panache Conseil

Du LHCOPN au LHCONE

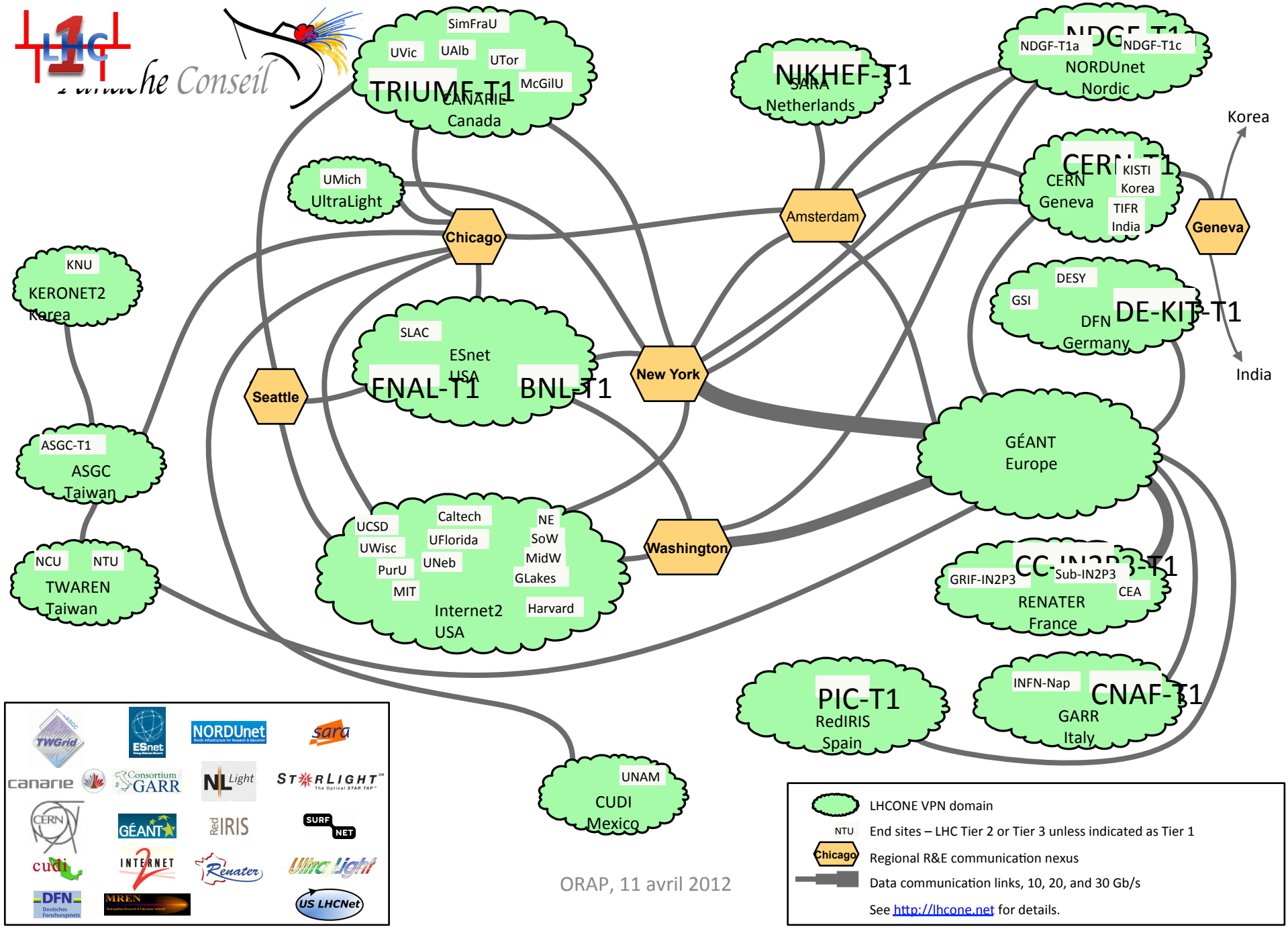
- Implication pour la France: VPN-L3 pour tous les Tier-2, possible grâce au maillage existant de RENATER; seconde adduction LHCONE-GEANT sur le NR de Genève; Adduction secondaire sur LHCONE-GEANT à Paris (en plus de l'interconnexion GEANT standard)
- La capacité de LHCONE est déjà du même niveau que celle de LHCOPN et continuera de croître!

Du LHCOPN au LHCONE

- Exemple de trafic, après <1 mois de MES:



LHCONE: A global infrastructure for the LHC Tier1 data center – Tier 2 analysis center connectivity



ORAP, 11 avril 2012

Panache Conseil

Conclusion pour LHC

- Qualité du dialogue entre NREN et HEP
- Bénéfice à postériori des choix d'architecture faits par les NRENs (réseaux hybrides maillés)
- Connaissance (même évolutive) de la matrice des flux entre nœuds
- La partie réseau du WLCG est réputée aujourd'hui pour être la plus fiable et la plus robuste!
- Avantage du modèle financier français qui permet des évolutions majeures de l'architecture sans bloquer sur des contraintes financières à court terme

Panache Conseil

Autres communautés

- Radio-astronomie
  - VLBI (Very Long Baseline Interferometry):
    - Les images du ciel prises par des antennes distantes sont corrélées pour produire une image globale correspondant à celle d'une antenne de taille égale à la surface de répartition des antennes.
    - Les données de chaque antenne étaient stockées sur bandes magnétiques, et rassemblées sur un corrélateur situé en Hollande.
    - 2 mois ("shipping") environ entre les prises de vues et la production de l'image globale
    - Le "shipping" a été remplacé par des liaisons GE: Le délai est réduit à 2 heures; Cela permet une autre approche de la radioastronomie!
    - Les télescopes attendent la mise en place de liaisons 10G; La corrélation peut se faire à l'échelon mondial!

Autres communautés

- Radio-astronomie
  - SKA (Square Kilimeter Array):
    - Prochaine génération de radio-télescope (Afrique du Sud ou Australie): MES: 2020-2025
    - Prochain défi des technologies de transmission:
      - Bande passante entre corrélateur et centre de traitement des données: 500 Tb/s
      - Distribution des données dans le monde entier à partir du site central: 100 Gb/s
      - La technologie n'est pas encore prête mais devra l'être pour le démarrage

Panache Conseil

Autres communautés

- Biologie
  - Depuis le milieu des années 90, les sciences de la vie annoncent qu'elles supplanteront les autres disciplines pour les volume des données et les puissances de calcul nécessaires.
  - La situation est en passe de devenir réalité aujourd'hui, grâce aux avancées de la génomique et de la protéomique!
  - Expérimentation en cours pour installer un circuit 100G entre Evry (IdGv) et Bruyère-le-chatel (CCRT/TGCC) pour décharger une batterie de séquenseurs sur les espaces de stockage!

Panache Conseil

Autres communautés

- HPC centralisé
  - Un supercalculateur (≥ Pflops) doit être considéré comme une source de données comparable au TGCC!
  - Les utilisateurs de ces données ne sont pas placés de façon déterministe comme les physiciens des particules!
  - La machine de calcul doit donc avoir ses propres moyens de stockage ou être capable de transférer ses données **aux utilisateurs, où qu'ils se trouvent, et éventuellement en temps réel pour les séries temporelles!**

Panache Conseil

Autres communautés

- HPC multi-sites DEISA
  - Mise en réseau au niveau européen des machines Tier-1
  - Impossible de connaître la matrice de trafics nécessaires entre centres Tier-1
  - Faute d'un cœur de réseau dédié pour DEISA sur GEANT, l'interconnexion s'est traduite par une topologie de circuits 10G en étoile, autour d'un commutateur à Francfort (4 sites Tier-1 en Allemagne, 1-2 site par pays pour les autres).

Panache Conseil

Autres communautés

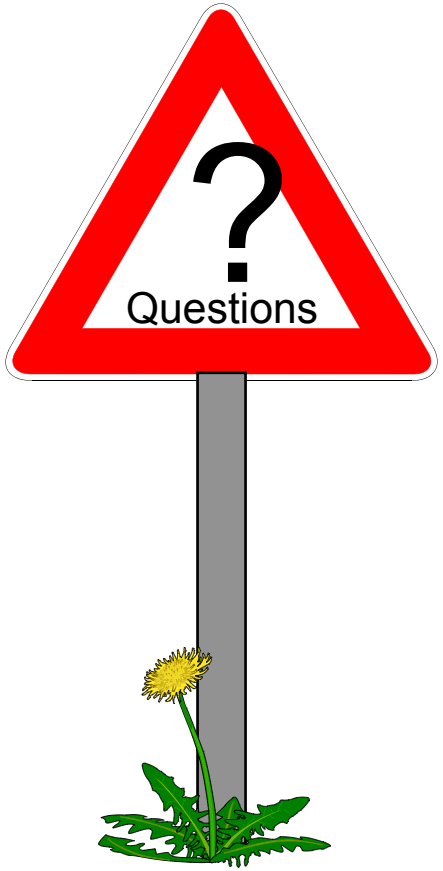
- HPC multi-sites PRACE
  - Projet d'interconnexion 100G entre Julich et Bruyère
  - La topologie nécessaire n'est pas encore identifiée si le nombre de Tier-0 est supérieur à 2!
  - En attendant, PRACE pourra disposer d'un cœur 100G entre 2 nœuds (architecture de la première version de TERAGRID aux US).
  - Possibilité de faire beaucoup plus performant si le dialogue NREN/HPC est approfondi…

Panache Conseil

Conclusions

- L'intérêt des réseaux hybrides est de séparer les problèmes d'infrastructure (NREN) des contraintes d'architecture (NREN + Utilisateurs)
- Nécessité d'un dialogue approfondi avec les communautés utilisatrices
- Nécessité pour les utilisateurs de connaitre suffisamment les matrices de flux de données pour définir l'architecture à mettre en place
- La connaissance de la taille de l'interface d'accès est insuffisante pour optimiser le réseau et son coût!
- Le modèle financier du NREN fera les choses possibles ou non!

FIN DE LA PRESENTATION



Et place aux questions!