Dominique Boutigny

L'infrastructure des salles informatiques

6 octobre 2009

ANGD Serveurs de calcul

Le CC-IN2P3 au CNRS

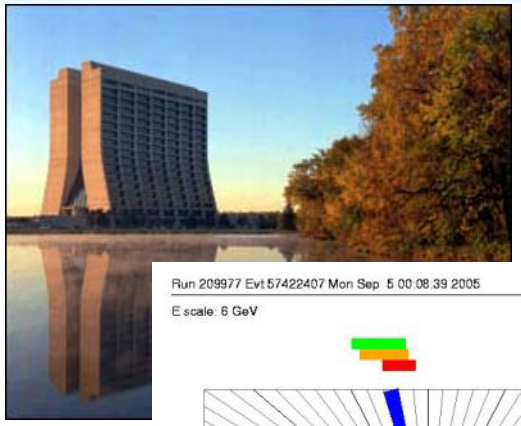
Institut National de Physique Nucléaire et de Physique des Particules

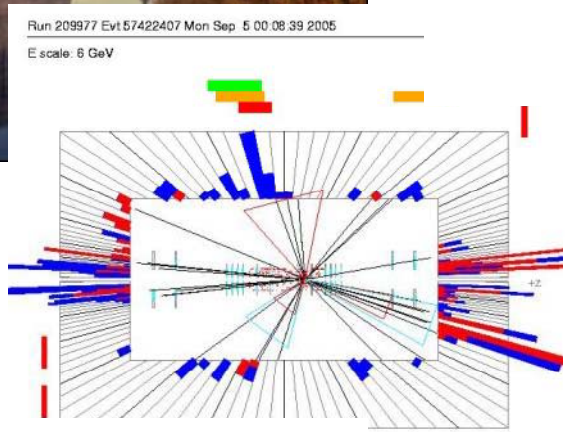
Le CC-IN2P3 centralise les moyens de calcul "lourds" de la communauté de
- Physique nucléaire
- Physique des particules
- Physique des astroparticules

Centre de calcul dédié

DSM Irfu Commissariat à l'Energie Atomique

Mission du CC-IN2P3





Expériences de physique

- Physique Nucléaire
- Physique des particules
- Astro-particules

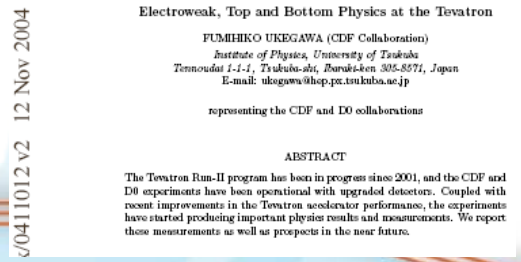
Masse de données

101000 100111 0001001010001
00011101 100010001111 00010
101000 100111 0001001010001
00011101 100010001111 00010
101000 100111 110001 111010
0001001010001101000 100111
0001001010001 00011101 1110
100010001111 00010 101000 00
11 0001100111 0001001010001
00011101 100010001111 00010
101000 100111 0001001010001

Recherche fondamentale

Publications

FERMILAB-CONF-
CDF/PUB/CDF/PUBI
November



0411012 v2 12 Nov 2004

Electroweak, Top and Bottom Physics at the Tevatron

FUMIHIKO UKEGAWA (CDF Collaboration)
Institute of Physics, University of Tsukuba
Tennoudai 1-1-1, Tsukuba-shi, Ibaraki-ken 305-8571, Japan
E-mail: ukegawa@hep.px.tsukuba.ac.jp

representing the CDF and D0 collaborations

ABSTRACT

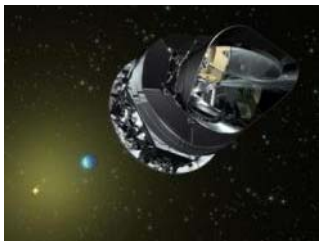
The Tevatron Run-II program has been in progress since 2001, and the CDF and D0 experiments have been operational with upgraded detectors. Coupled with recent improvements in the Tevatron accelerator performance, the experiments have started producing important physics results and measurements. We report these measurements as well as prospects in the near future.

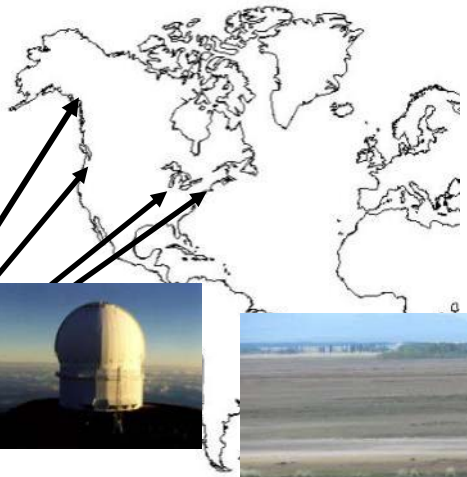
Analyse des résultats

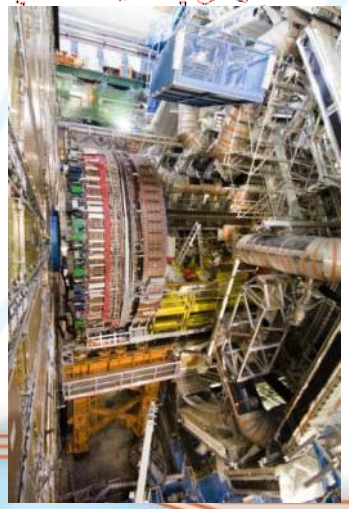
Traitement des données

Une dimension internationale

Le CC-IN2P3 fait partie d'un réseau mondial de grands centres de calcul pour la physique des hautes énergies

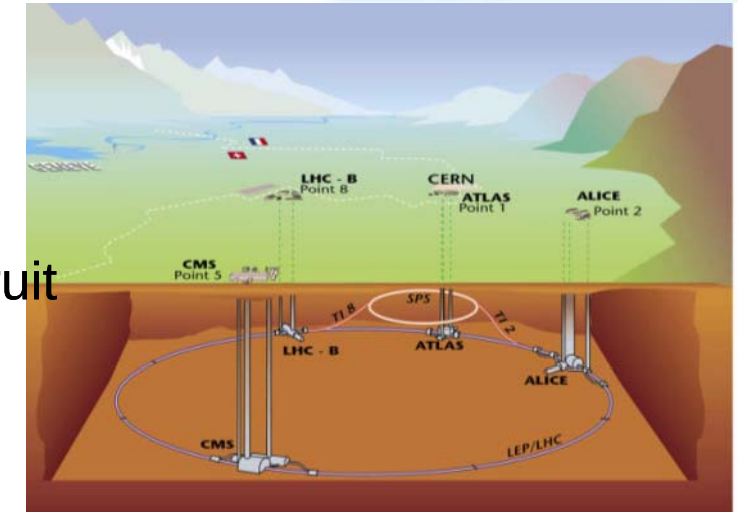






Le LHC

On peut utiliser beaucoup de superlatifs pour qualifier le LHC



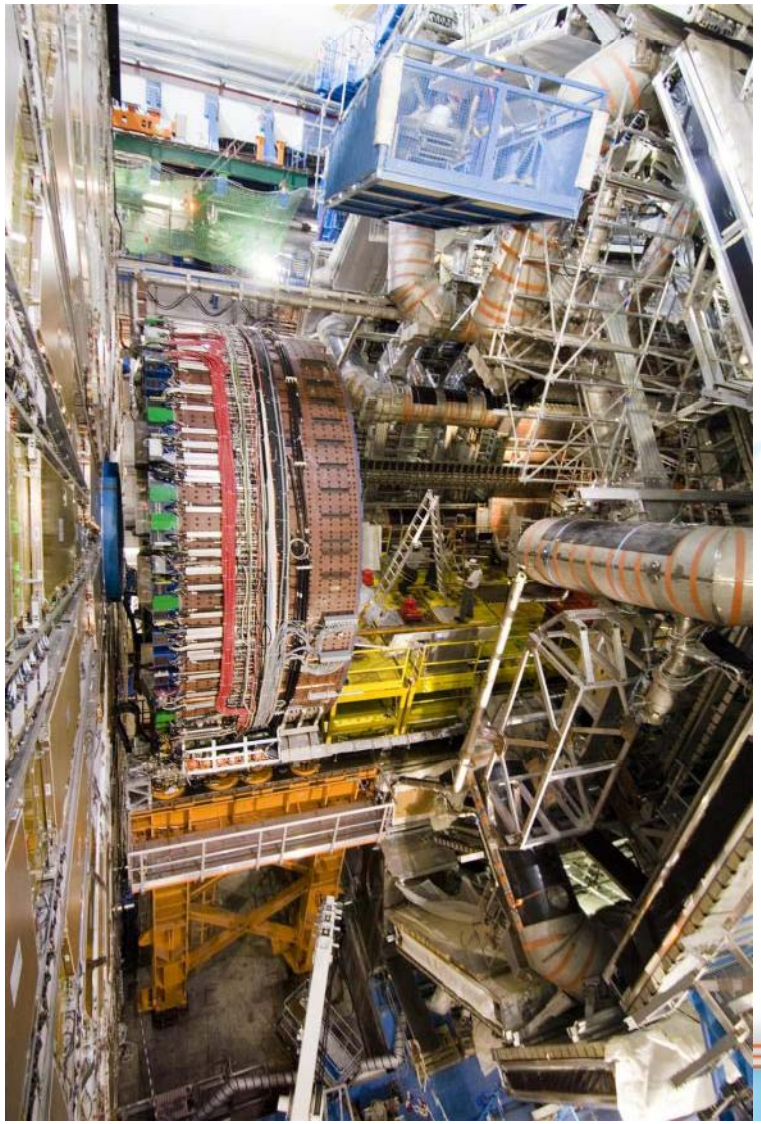
- ✓ Plus grand accélérateur du monde (27 km)
- ✓ L'un des plus grand instrument jamais construit
- ✓ Plus grande installation cryogénique
- ✓ Plus froid que l'espace interstellaire
- ✓ Mais des collisions 100 000 fois plus chaude que le cœur du soleil

… et ainsi de suite…

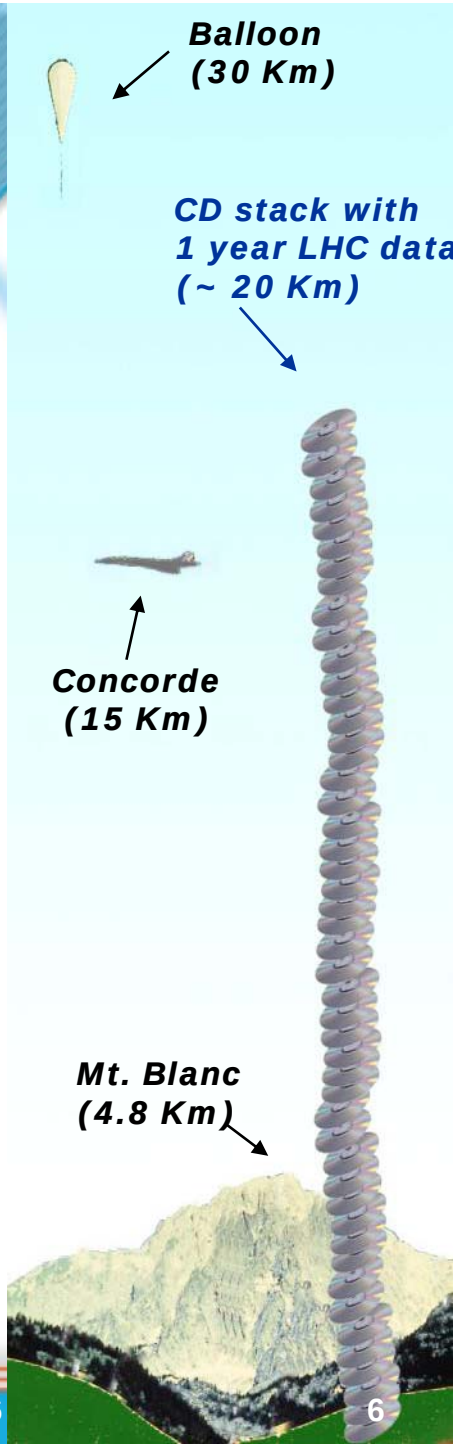
4 expériences / collaborations internationales

ALICE – ATLAS – CMS - LHCb

Le LHC – Un énorme générateur de données

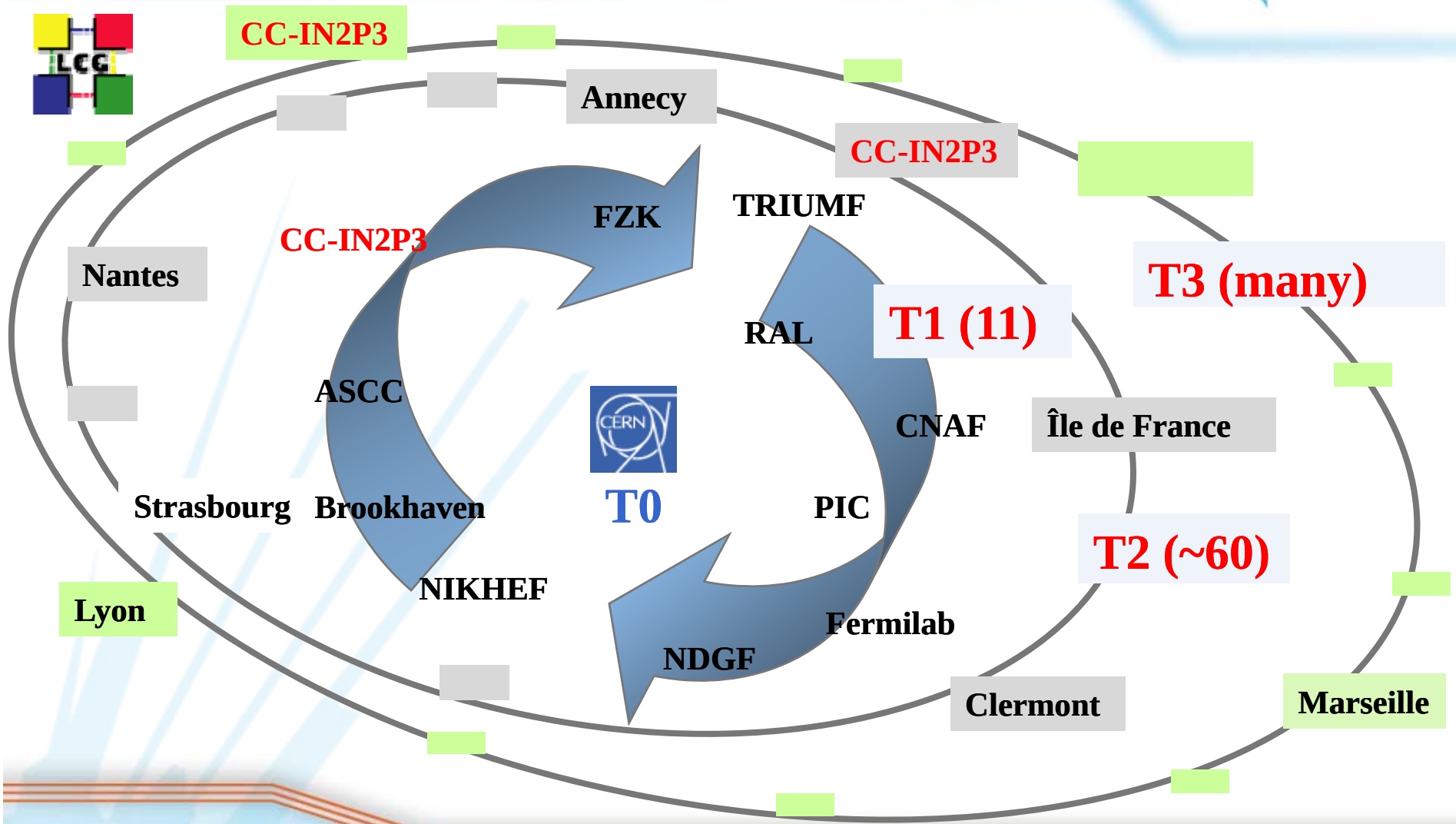


15 pétaoctets de données chaque année

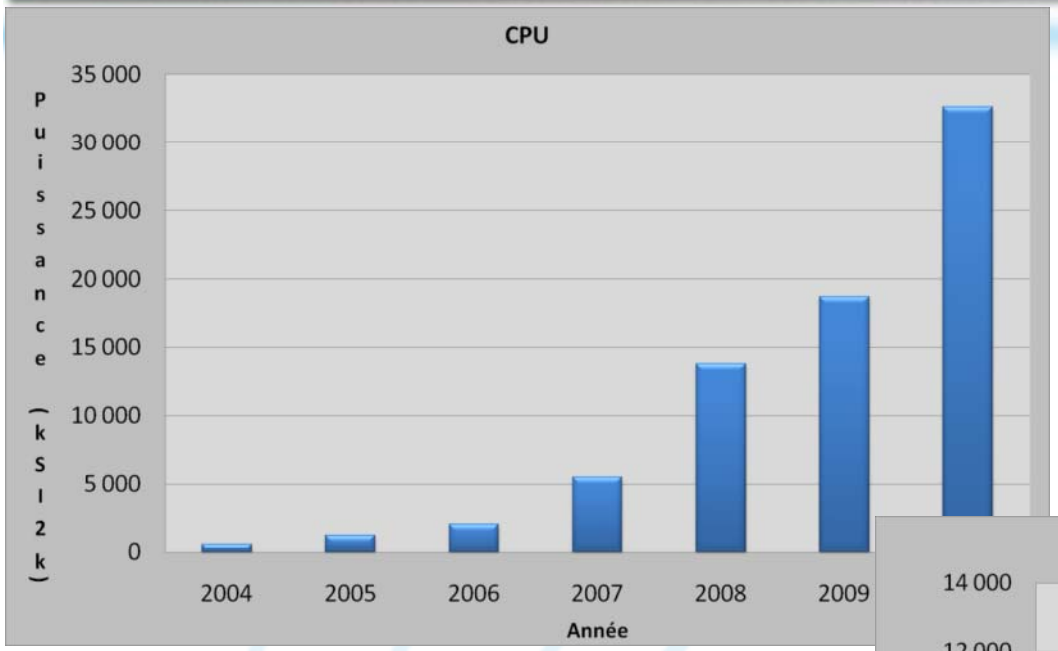


6

Une architecture de Grille globale

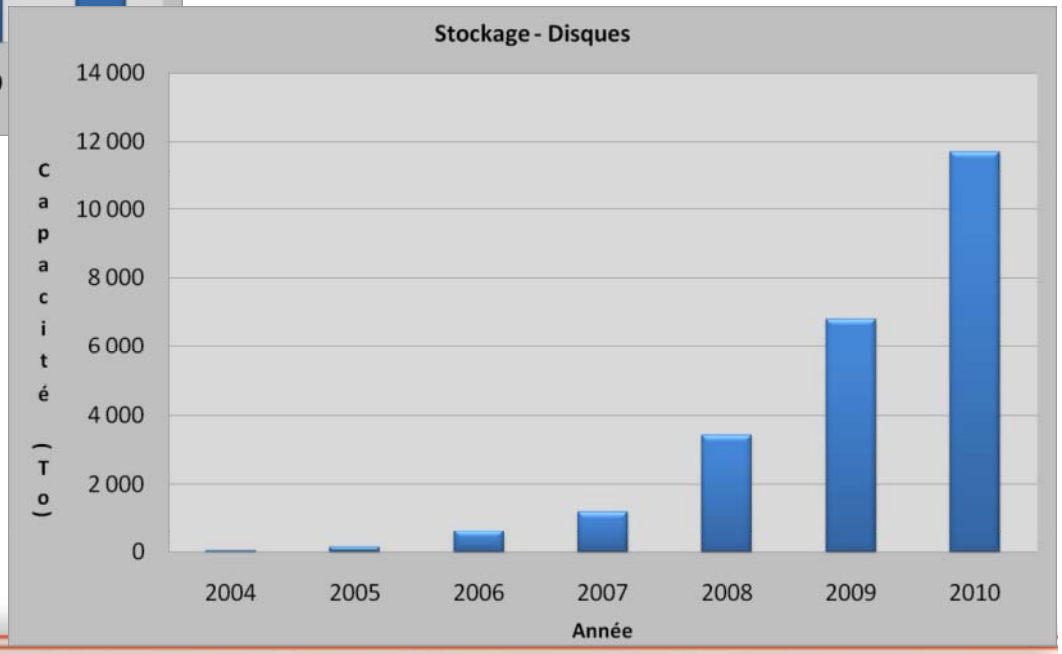


Évolution des ressources au CC-IN2P3



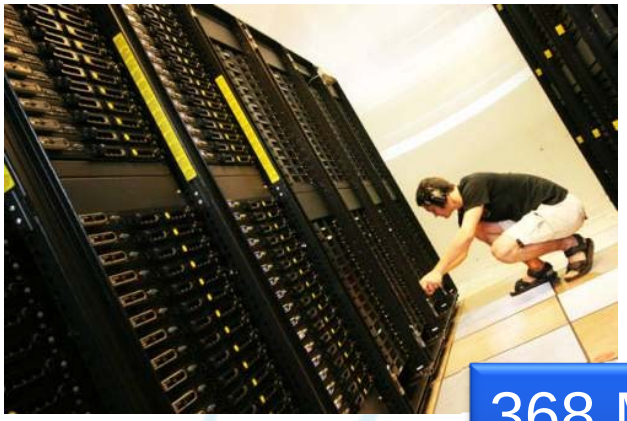
Noter l'unité de puissance CPU en kSI2k

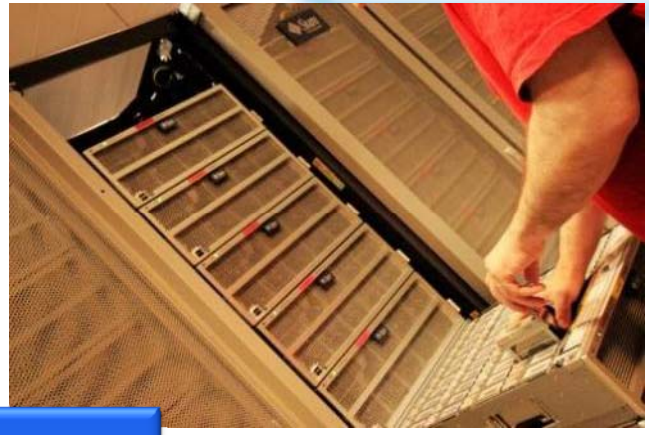
+ 3 robots de stockage de masse (cassettes) 30 Po



Aujourd'hui le CC-IN2P3 c'est …

1398 machines Linux





368 Machines Solaris
Dont: 302 Thumpers / Thor

- 14 496 disques
- 2416 RAMs
- 604 CPUs

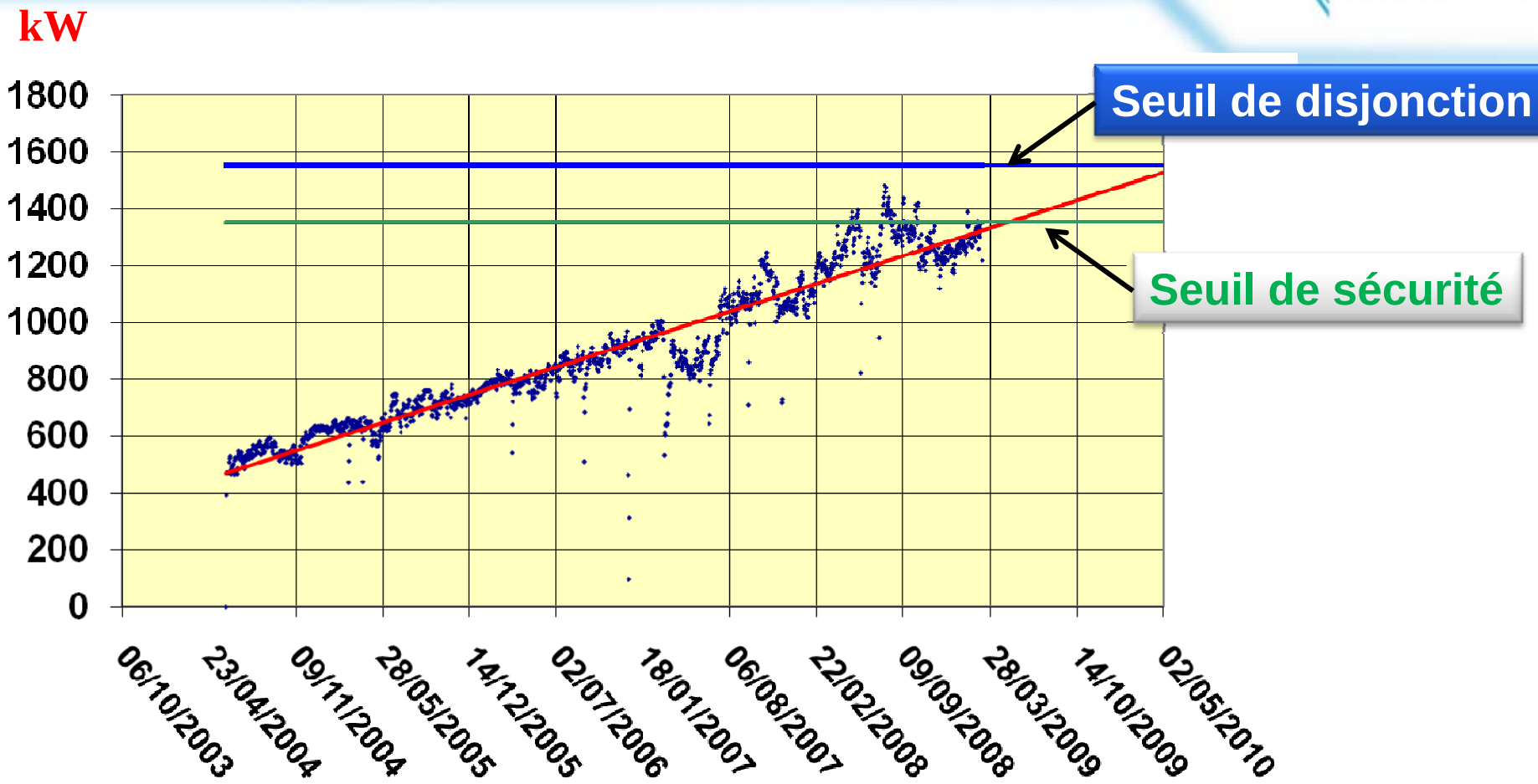
100 Machines AIX

Le problème de l'infrastructure

Le problème est d'accommoder une augmentation exponentielle de la puissance informatique avec une salle machine conçue 20 ans plus tôt

Le problème est le même lorsque l'on veut installer un cluster moderne (même de taille modeste) dans des locaux non adaptés

Évolution de la puissance électrique



Accroissement de 500W / jour !

Budget élec. 2009: 600 k€

Consommation Elec. / Puissance CPU

Un centre de calcul ne peut plus raisonner seulement en terme de puissance CPU ou d'espace de stockage

Puissance CPU

Puissance électrique

Capacité de stockage

Encombrement

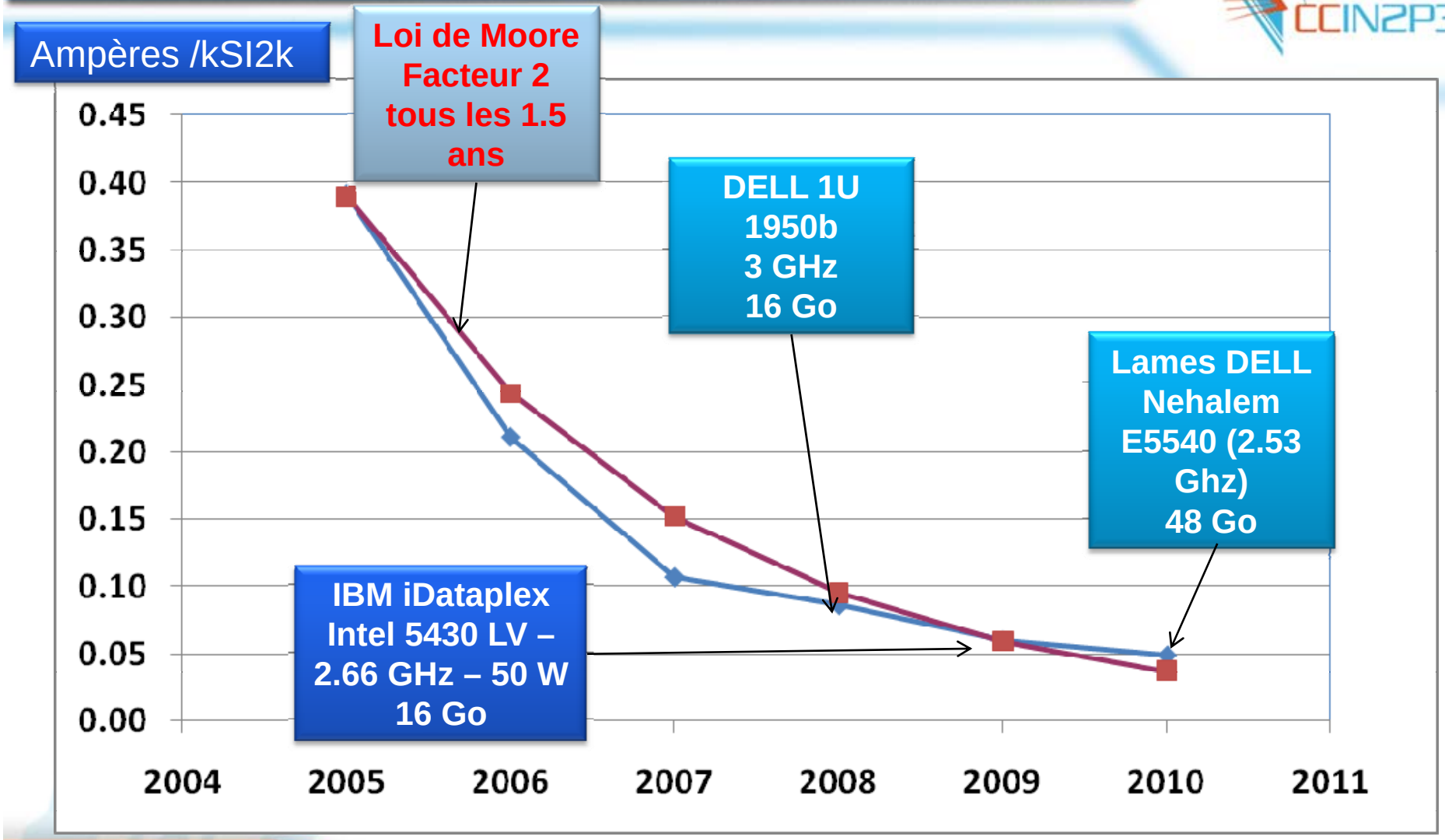
Facilité de mise en œuvre et d'exploitation

Ampères ou W / kSI2k
Ampères ou W / To

Respect de l'environnement

Élément prépondérant du cahier des charges

Consommation des CPU



Consommation des disques

La consommation du stockage sur disque est loin d'être négligeable

1 SUN X4500: ~1000 W ➔ disques de 500 Go

17 To utile / serveur ➔ **~60W / To**

Nouvelle génération X4540 ➔ disques de 1 To

On gagne un facteur 2 à chaque changement de technologie

Optimisation de la consommation des serveurs de stockage ?

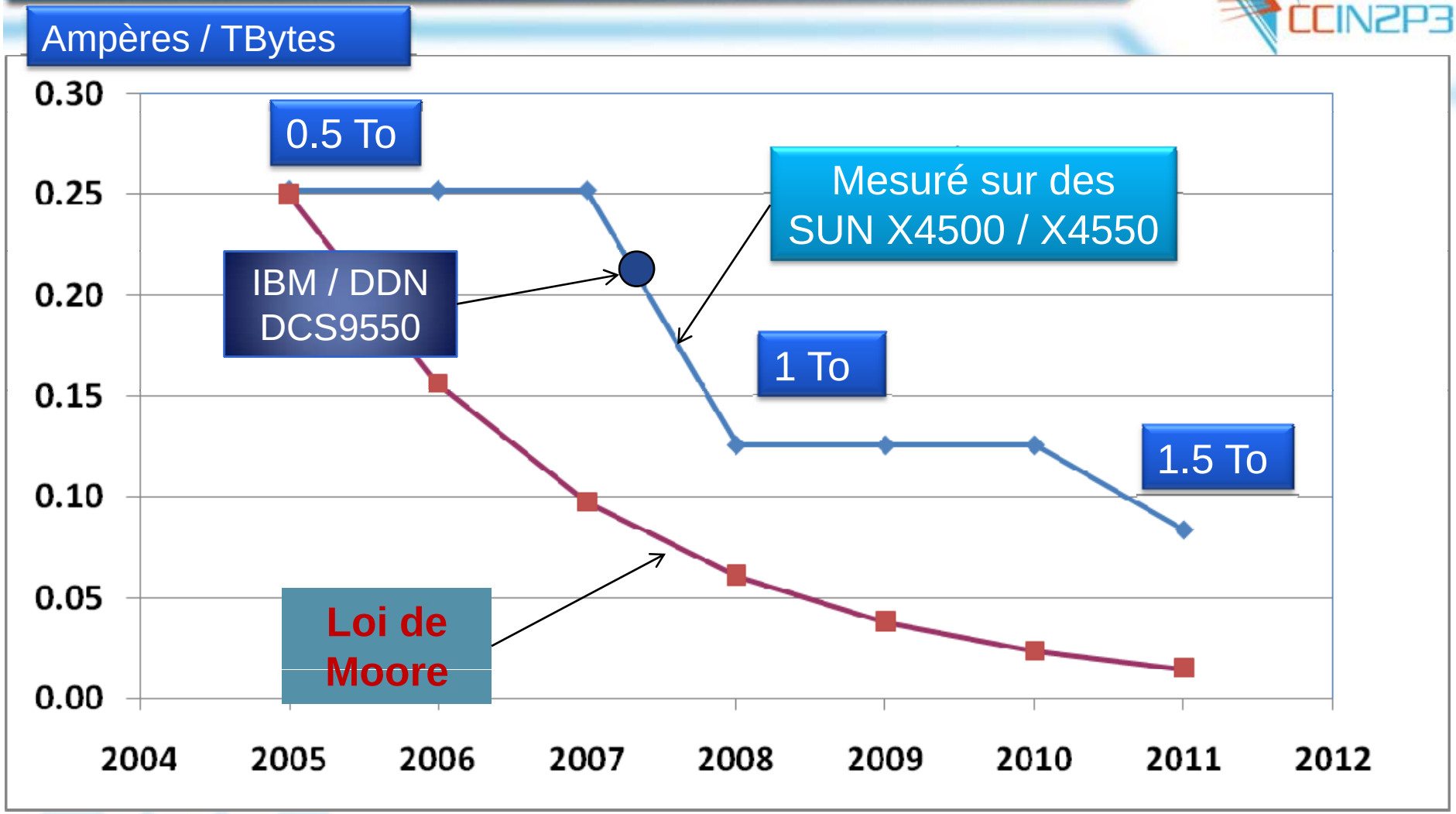
Ajout de cache SSD

Modification des drivers pour arrêter les disques inactifs

➔ Technologie MAID (Massive array of idle disks)

Voir par exemple: http://www.green-bytes.com/ZFSplusnew.html

Consommation des serveurs de disques

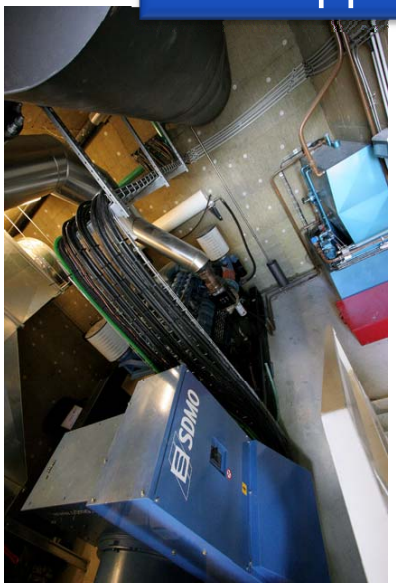


kW – kVA – Cos Φ

Il faut faire très attention aux unités que l'on utilise

Un serveur utilise des Watts

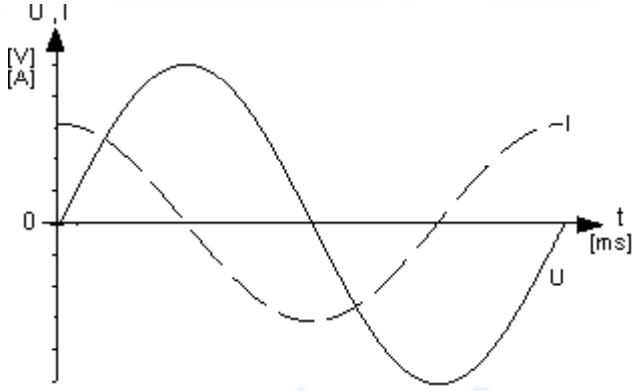
Un groupe électrogène développe des kW



Un disjoncteur est calibré en Ampères

Attention les constructeurs utilisent les unités qui les arrangent !

kW – kVA – Cos Φ



Les éléments capacitifs ou inductifs introduisent un déphasage Φ entre le courant (A) et la tension (U)

$$P = U \times I \cos \Phi$$

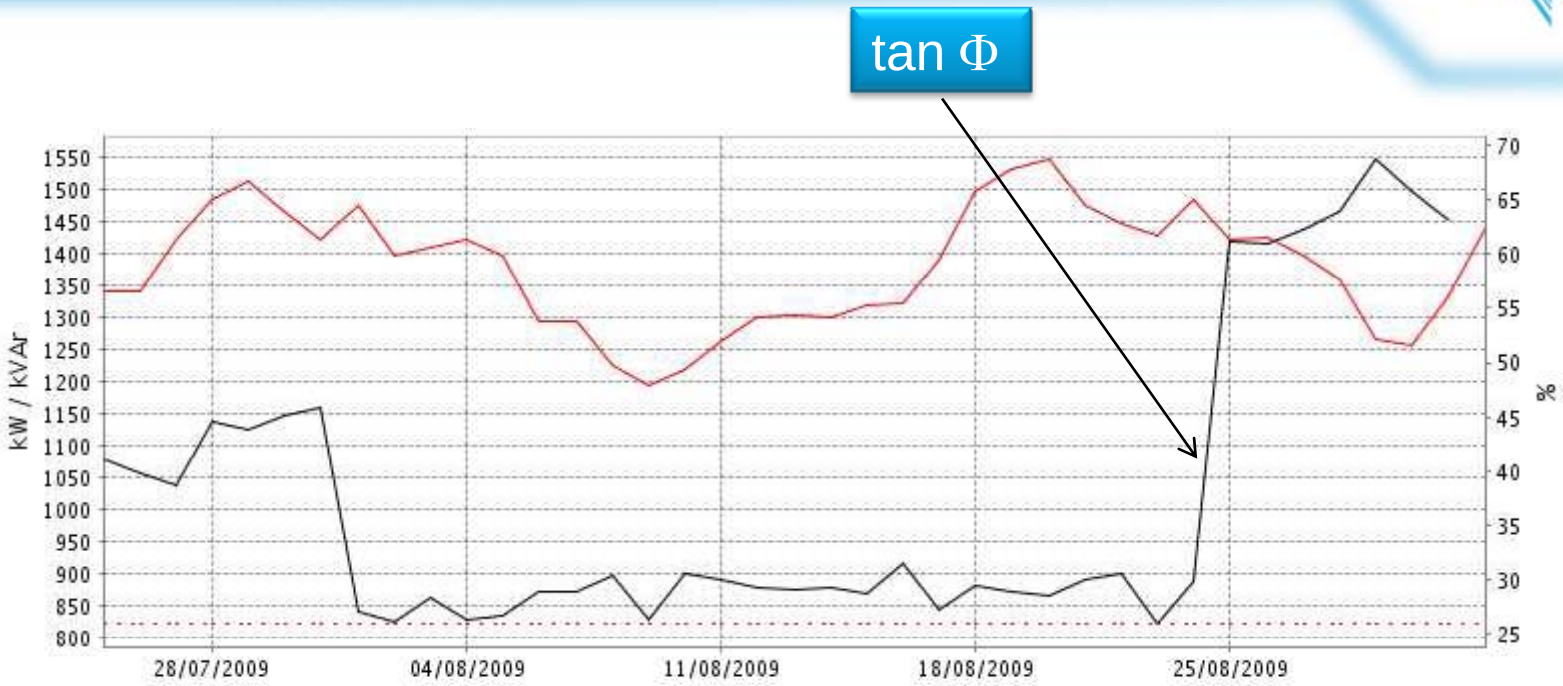
Dans les salles informatiques le déphasage est essentiellement inductif (bobines)

Si Φ est important (cos Φ est petit) alors pour une même puissance il faut fournir plus d'Ampères ➔ risque de disjonction

Un mauvais cos Φ est pénalisé par EDF (0.8 ?)

➔ Notions de puissances active et réactive

kW – kVA – Cos Φ



Il y a tout intérêt à monitorer le déphasage
EDF fournit le service Adviso qui permet de suivre en ligne les consommations et le déphasage

Et si le déphasage est trop grand ?

➔ Il faut installer un compensateur de cos Φ

Compensation dynamique grâce a des batteries de condensateurs et / où d'inductance

Celui-ci doit être correctement dimensionné et placé au bon endroit !

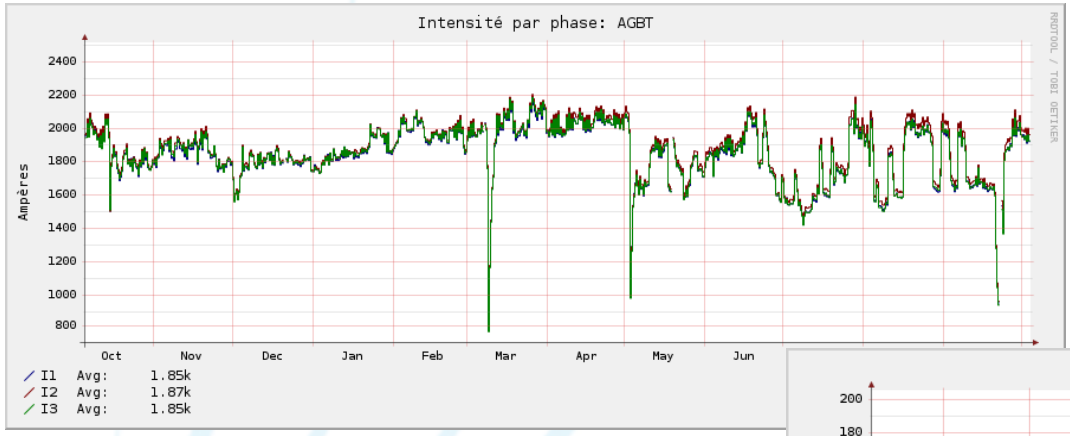
L'importance du monitoring

Un bon monitoring des consommations électriques est crucial

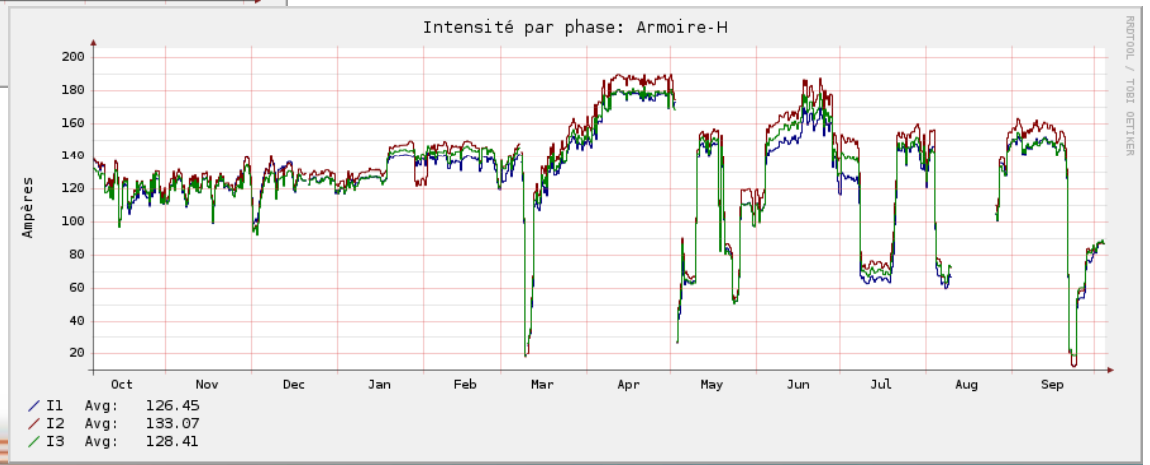
➔Armoires électriques équipées de tores

➔PDA adressables

➔Châssis "intelligents"



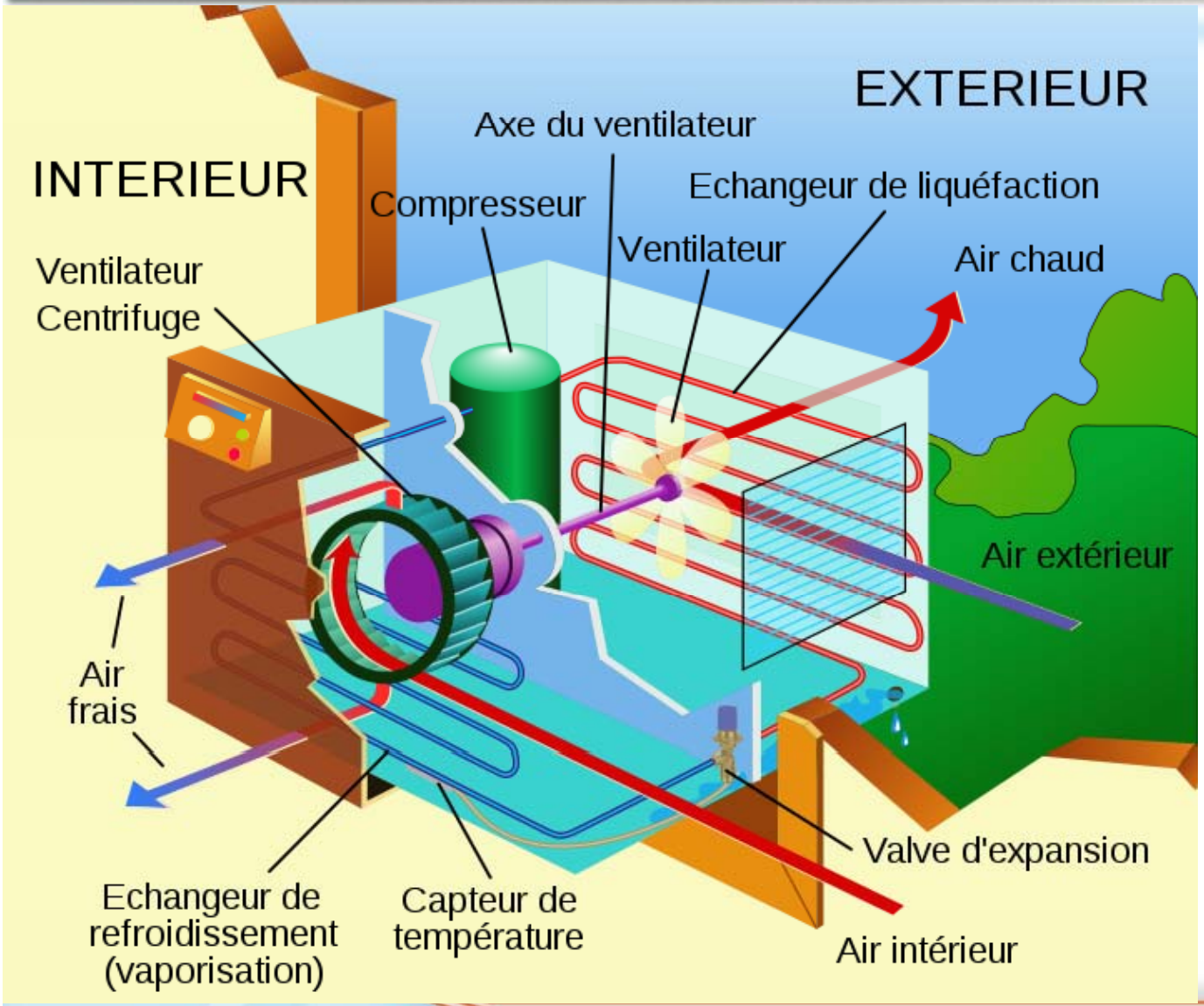
Idéalement on doit être capable de connaître la consommation en temps réel de chaque équipement



Contrôle thermique des installations

Un contrôle annuel des armoires électriques avec une caméra thermique est **très** utile ➔ détection des mauvaisesconnexions

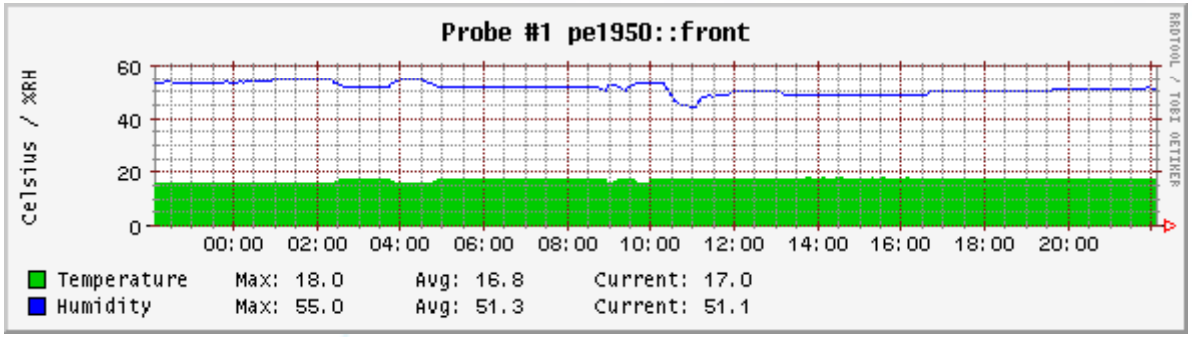
Principe du climatiseur

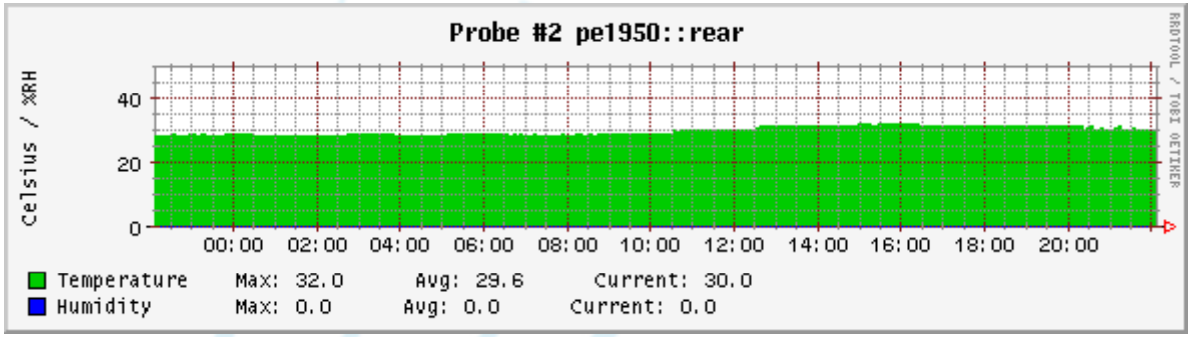


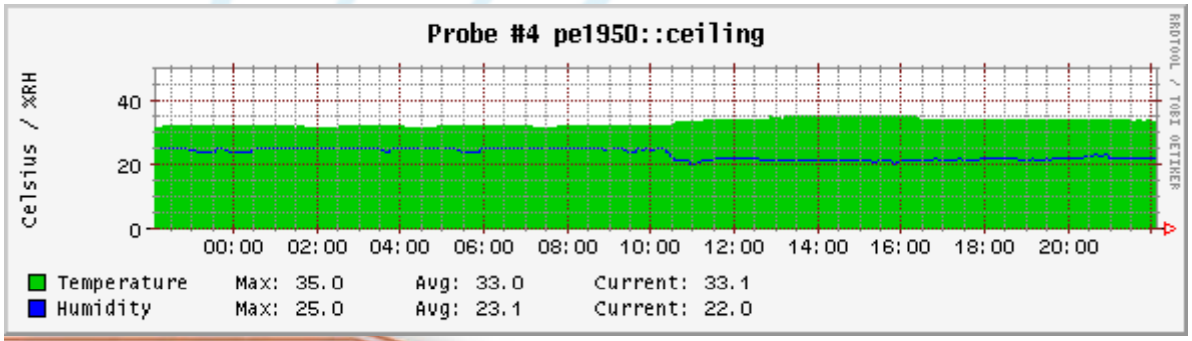
Source: Wikipedia


05.10.2009 16:36

Le problème du refroidissement







A moins d'avoir un système aéraulique complexe, Le refroidissement par air suppose de traiter le volume complet de la salle informatique

➔ Peu efficace

Notion de PUE

Source: F. Berthoud
Journée mésocentres

PUE : Power Usage Efficiency

Ratio entre le dépense énergétique totale d'un datacenter et l'énergie effectivement consommée par le matériel informatique. Compris entre 1 et 3 ou 4 ou plus … (DCiE : inverse du PUE (compris entre 0 et 100%))

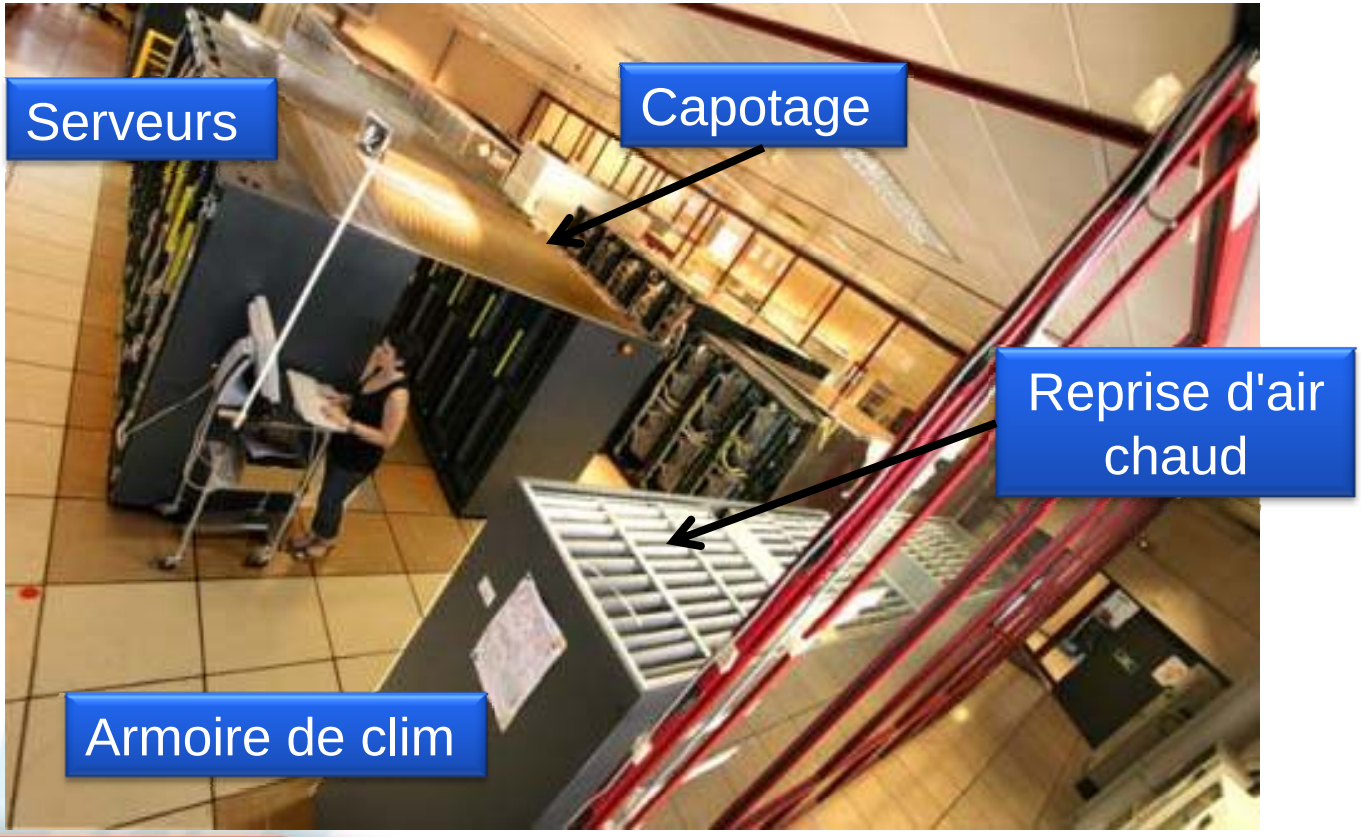
En moyenne, plus de 60% de l'énergie est consommée par le froid, le système électrique lui-même, l'éclairage etc. (PUE=2.5)

Attention au PUE, c'est un indicateur parmi d'autres..

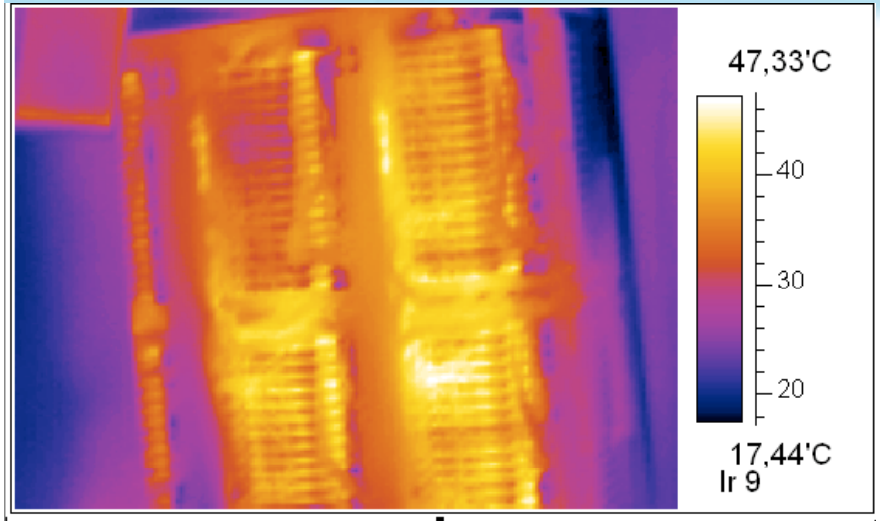
Exemple: L'alimentation électrique des serveurs fait partie du matériel informatique. Une alimentation inefficace ne dégrade pas le PUE, et pourtant…

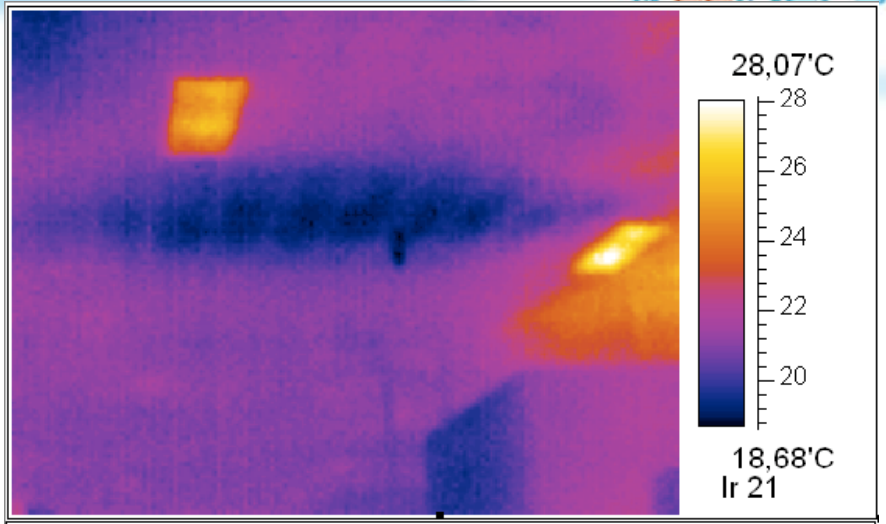
Refroidissement

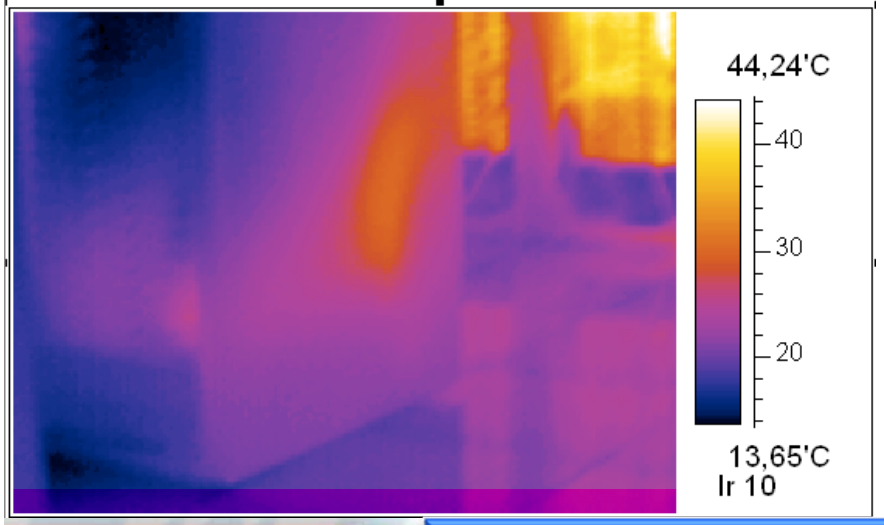
Organisation en allées chaudes / allées froides

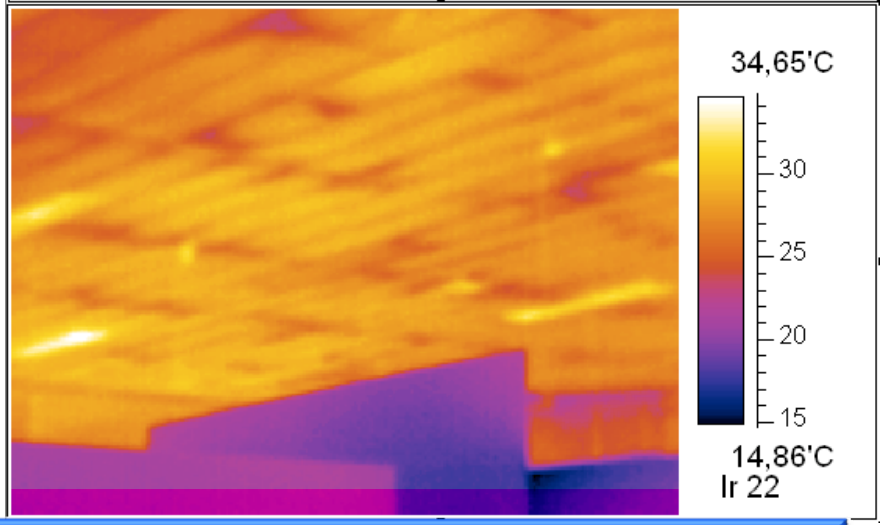


Un comportement non intuitif…





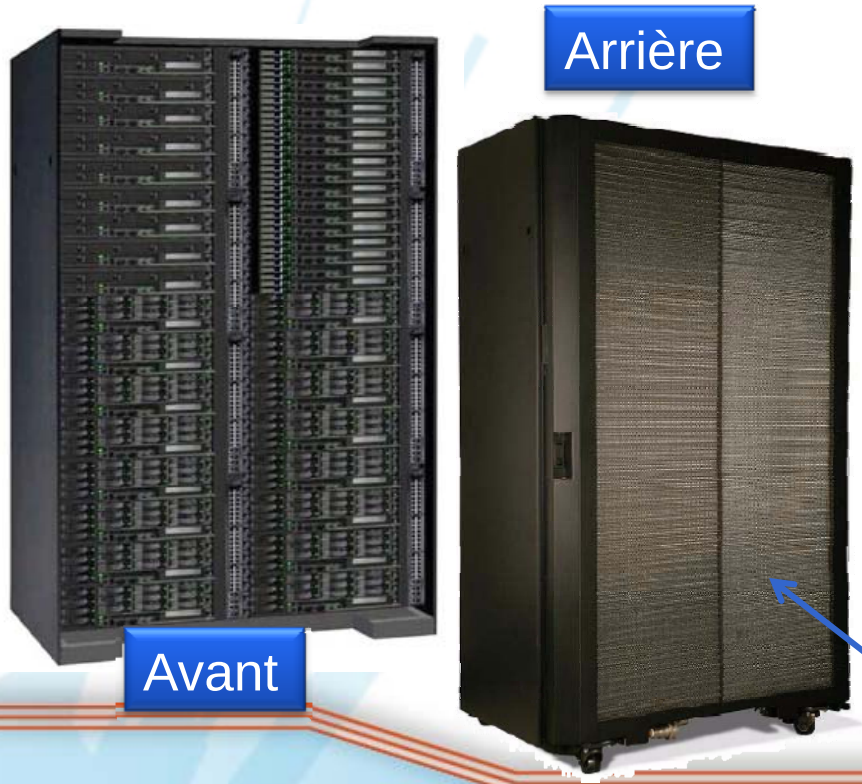




Un monitoring efficace est indispensable afin de comprendre en temps réel le fonctionnement de la salle

Comment survivre ?

Améliorer l'efficacité du refroidissement
Air ➔ Eau



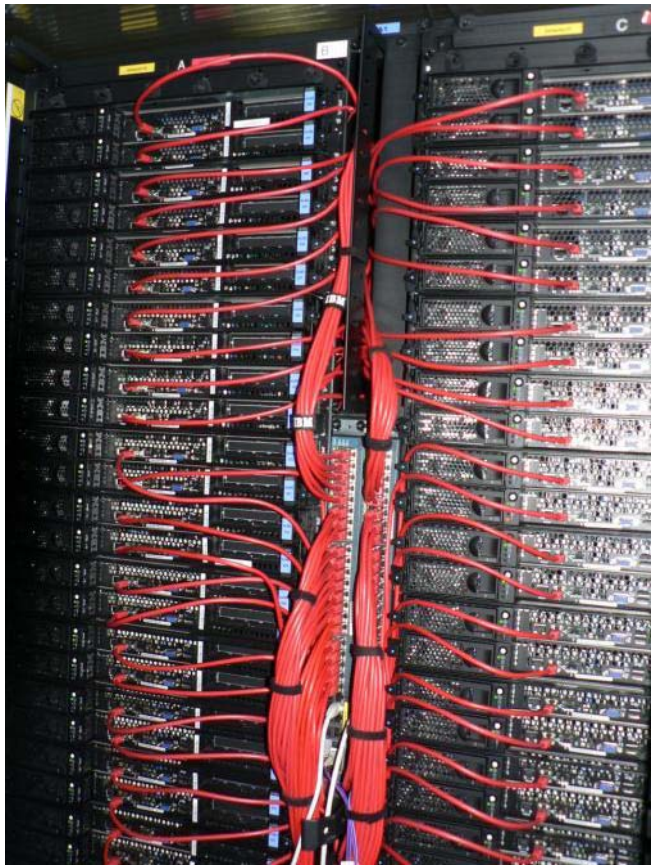
Système IBM i-dataplex
412 serveurs (5 racks) – Intel 5430 LV – 2.66 GHz – 50 W
85 kW total – ~pas de chaleur en dehors des racks

Aussi efficace que des racks entièrement fermés

Moins cher que des Blades fin 2008

iDataplex et portes froides

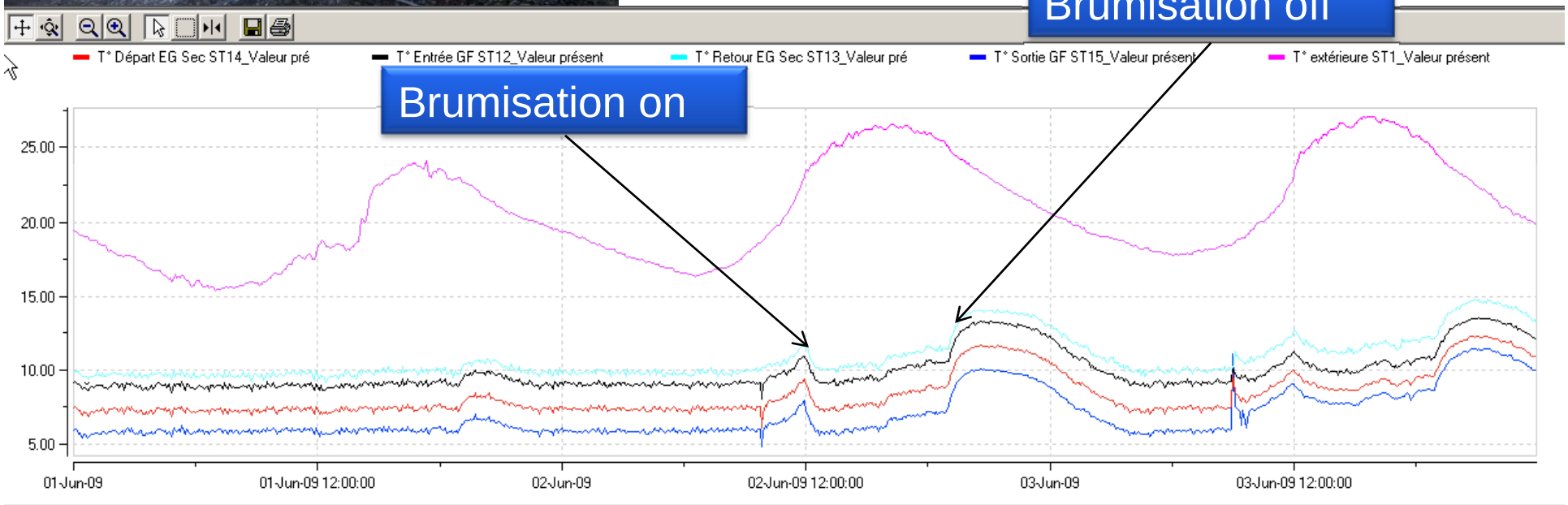
Fonctionne avec de l'eau à 18° ➔ pas de condensation



Nécessite l'installation d'un échangeur eau glacée – eau à 18°

Quelques astuces!

Brumisateurs d'eau installés sous les aéroréfrigérants



Refroidissement adiabatique

http://www.almeco.eu/bq/produits-services/systemes-a-haute-pression.html

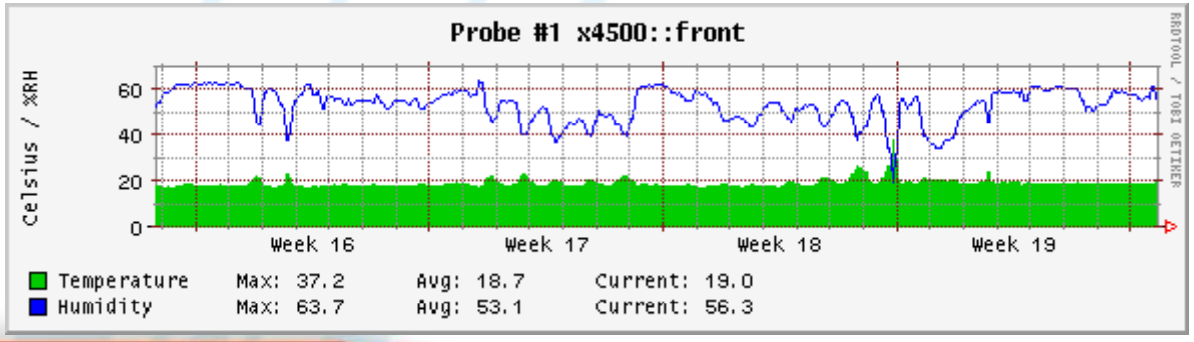
L'efficacité dépend de l'humidité relative (fonctionne bien par temps sec)

Attention, cette technique est à distinguer des tours de refroidissement qui sont soumises à une réglementation très stricte en raison des risques de légionelloses

Quand les choses dérapent…

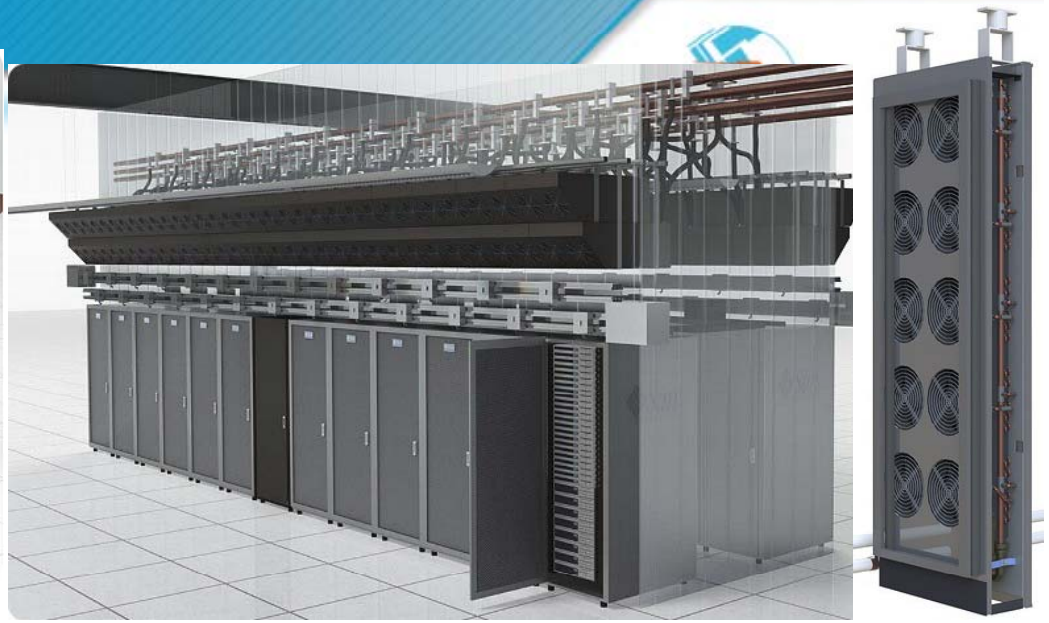
Lorsqu'on exploite une salle au maximum de sa capacité une panne sur un groupe froid peut rapidement dégénérer

➔ Effet boule de neige sur les groupes froids



La température a atteint plus de 60° dans certaines zones à l'arrière des serveurs

L'étape d'après…



Confinement des allées chaudes et refroidissement intégré aux racks

Racks RITTAL
Coût: 110 k€ pour 8 racks travaux compris

La fin des faux plancher

Il faut encore vaincre la résistance des bureaux d'étude !

Avec les techniques de racks refroidis avec un circuit d'eau glacée, les faux planchers sont devenus inutiles

Passage des câbles et des tuyaux par le haut

➔ Visibilité – Clarté des cheminements – Facilités de branchement – Résistance du sol

Remarques

Il y a une limite à la densité kW / rack

Actuellement cette limite se situe autour de 30 kW / rack. Au-delà la densité effective des équipements informatiques diminue en raison de l'augmentation du nombre de dispositif de climatisation

L'abandon du faux plancher rend encore plus important le fait de penser à l'avance l'urbanisation de la salle

- Positionnement des chemins de câbles
- Dimensionnement du réseau d'eau glacée

Le Free Cooling

Le free cooling est une technique qui permet de refroidir une salle informatique avec l'air extérieur lorsque celui-ci est suffisamment frais

Pourquoi climatiser en hiver lorsqu'il fait 5° dehors ? ou "quand il fait chaud j'ouvre la fenêtre" !

Les serveurs modernes sont plus robustes qu'avant et supporte des conditions de température et d'humidité moins contraignantes.

ASHRAE: 18° < T < 27° Humidité: > 5.5° Dew Point et < 60% RH

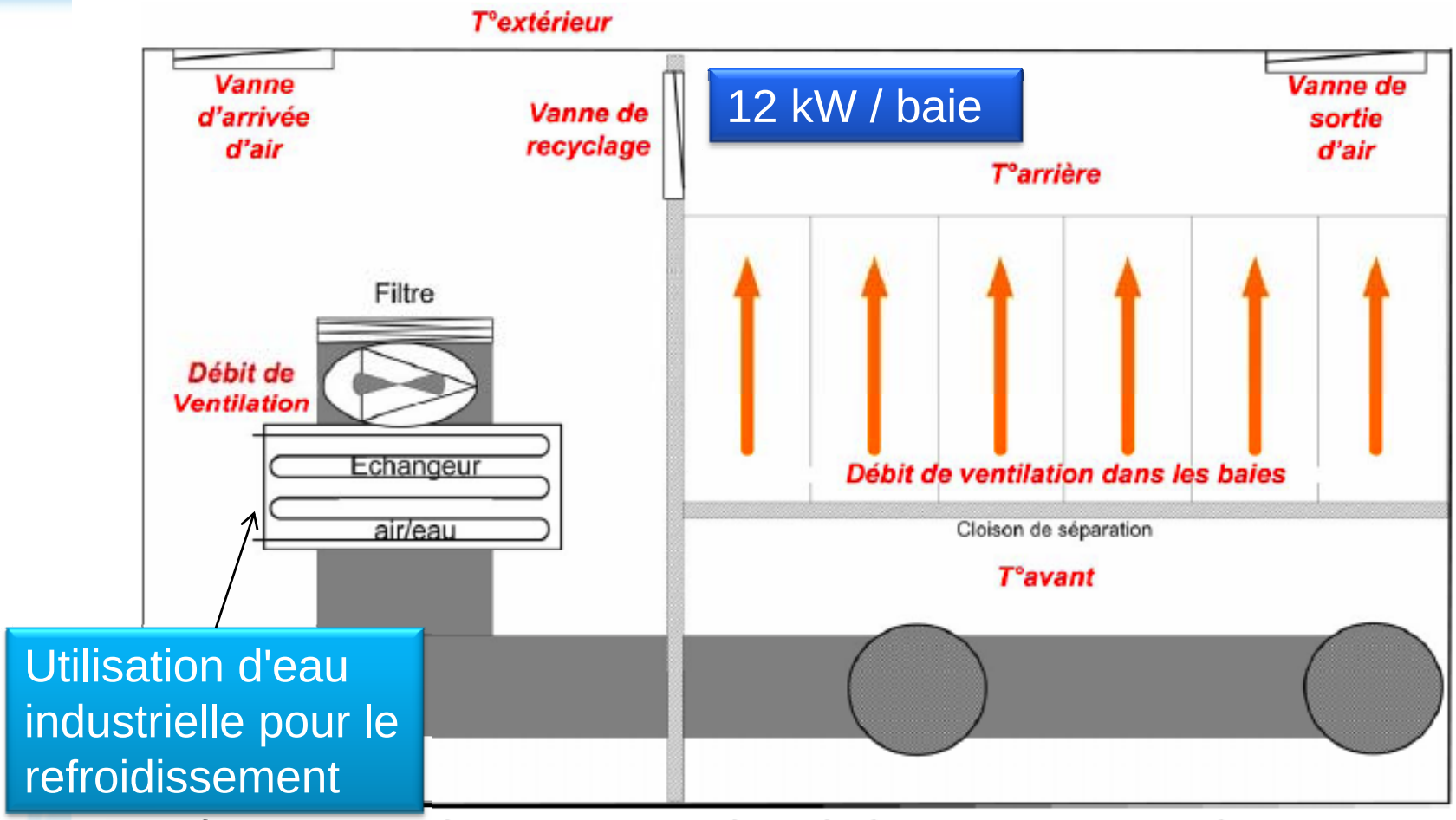
On peut dépasser ces valeurs pendant de courtes périodes

http://www.ashrae.org/

American Society of Heating, Refrigerating and Air-Conditioning Engineers

Projet ECOCLIM (LPSC)

Crédit: Bernard Boutherin



Le débit de ventilation d'air dépend de la puissance installée et est égal à la ventilation produite par les unités centrales des baies

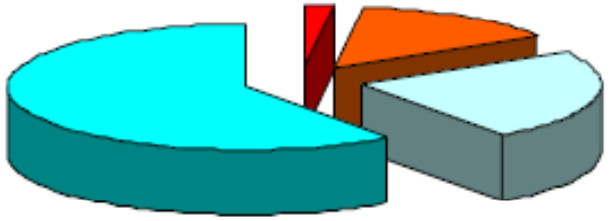
Jusqu'à 22 000 m^3/h de débit d'air

Projet ECOCLIM (LPSC)

Crédit: Bernard Boutherin

Régime canicule t°ext > 33°C 2% du temps
La consigne de température sera 25°C
Echangeur en fonction
t°consigne = 25°C régulation par débit d'eau
Vanne-recyclage ouverte = 1
Vanne-in fermée = 0
Vanne-out fermée = 0

Régime chaud 25° <t°ext <= 33°C 14% du temps
La consigne de température sera 25°C
Echangeur en fonction
t°consigne = 25°C régulation par débit d'eau
Vanne-recyclage fermée = 0
Vanne-in ouverte = 1
Vanne-out ouverte = 1



Régime froid t° ext <= 18°C 60% du temps
La consigne de température sera la consigne minimale soit 18°C,
Echangeur non utilisé
t°consigne = 18°C régulation par vanne recyclage
Vanne-recyclage ouverte = (18° - t°ext) x K1
Vanne-in ouverte = 1 – Vanne-recyclage
Vanne-out ouverte = Vanne-in

Régime normal 18° < t°ext <= 25°C 24% du temps
La consigne de température sera la température extérieure t°ext,
Echangeur non utilisé
t°consigne = t°ext pas de régulation
Vanne-recyclage fermée = 0
Vanne-in ouverte = 1
Vanne-out ouverte = 1

ECOCLIM – Bilan

Crédit: Bernard Boutherin

Coût:

- 40 k€ pour le système
- 20 k€ de travaux annexes

- Economie d'énergie
  - 84% du temps on économise 75% de la puissance de refroidissement
    33% de la puissance installée avec un COP de 3 pour une pompe à chaleur
    8% correspond à la puissance de ventilation nécessaire
  - Aujourd'hui avec 40KW installés on économise 10KW (84% du temps)
    Soit 75 000 KWh/an, (30 tonnes de CO2/an si on utilisait la filière fioul !)
  - Demain à pleine puissance ce sera :
    150 000 KWh/an d'économie, opération **remboursée** en 5 ans

Free Cooling

Ce concept peut être poussé très loin

Projet de datacenter à Stanford 6 modules de 6 MW refroidis par air
Un étage entier de collection d'air

Datacenter Microsoft à Dublin

PUE: 1.25

A terme: 22 MW

55 000 m²

Redondance des équipements

Lorsqu'on défini un centre de calcul il faut bien penser aux niveaux de disponibilité que l'on veut garantir pour les différents services

Par exemple:

	Level 3 CPU	Level 2 Stockage	Level 1 Haute dispo.
Refroidissement	N+1	N+1	2N
Electricité	N	N+1	2N
Alimentation	Single	Dual	Dual
Autonomie	< 10 minutes	72 heures	72 heures

Les redondances coûtent cher !

Quels équipements doivent être impérativement ondulés ?
Equilibrer les risques – Combien de coupure EDF par an ?

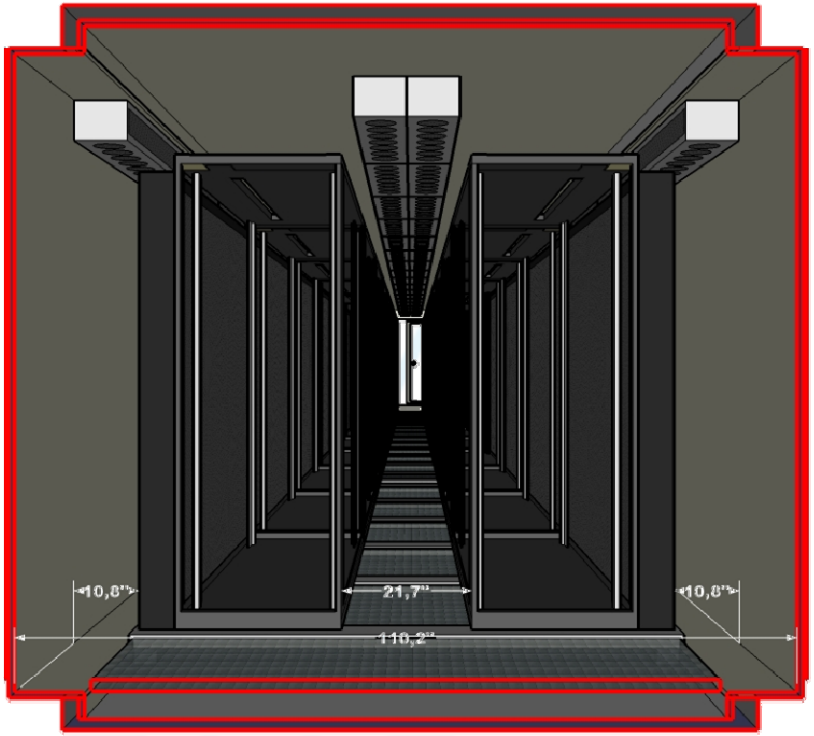
Alimentation sans coupure

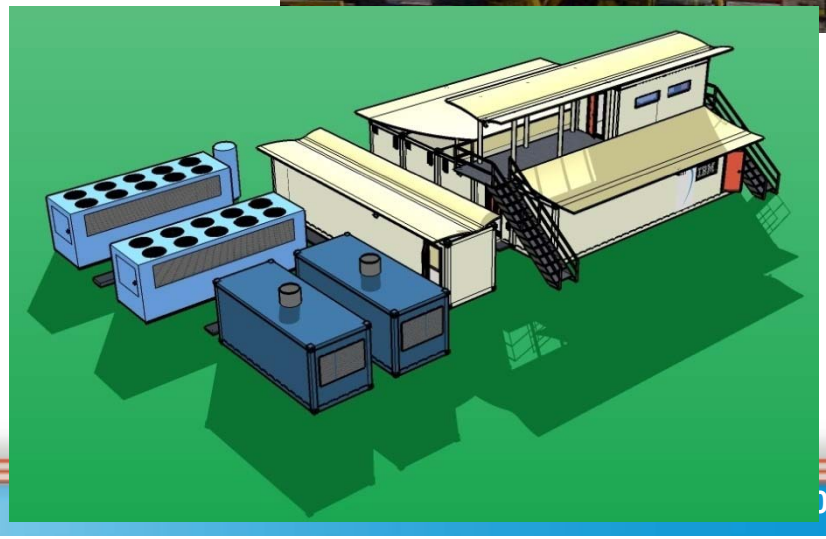
Des alternatives:

- Alimentation par 2 postes sources indépendants
- Groupes tournants (avec volant d'inertie)

Les containers

Les containers SUN Black Box ou IBM PMDC (Portable Modular Data Center) peuvent constituer des alternatives intéressantes à la construction d'une salle machine





Photos © IBM

Récupération de la chaleur

Une possibilité intéressante est de récupérer la chaleur de la salle informatique pour le chauffage

Le CC-IN2P3 est uniquement chauffé par sa salle machine

Inconvénients:

- Efficace surtout l'été !
- Eau relativement peu chaude (~55°)

Chauffage de piscine – de cantine etc…
Chauffage "basse température"

Législation ICPE

Installations Classées Protection de l'Environnement

Les normes ICPE sont très strictes
Géré par la DRIRE: Direction Régionale de l'Industrie, de la Recherche et de l'Environnement – Dépend de la préfecture

Puissance absorbée dans les groupes froids au Fréon > 500 kW
➔ Procédure d'autorisation

- Enquête publique
- 1 an de délai
- Opération du centre régie par un arrêté préfectoral

Déclenche une évaluation et des contraintes sur tous les risques environnementaux