Apport de l'intelligence au coeur des réseaux de communications pour de nouveaux protocoles et services dans l'Internet du futur

L. Lefèvre

RESAM Laboratory UCB – Action INRIA RESO

Ecole Normale Supérieure de Lyon, France

Laurent.Lefevre@inria.fr

Plan

- Protocoles dans l'Internet
- Une solution : Réseaux actifs
- Tamanoir
- Grille active
- Services actifs (Multicast, QoS)
- Projets : VTHD++, Etoile

Protocoles de l'Internet

- Tout le monde l'utilise :

IP

- Adressage : IPv4
- Tables de routage
- Non connecté
- Protocole ouvert
- Bientôt : IPv6

Protocoles

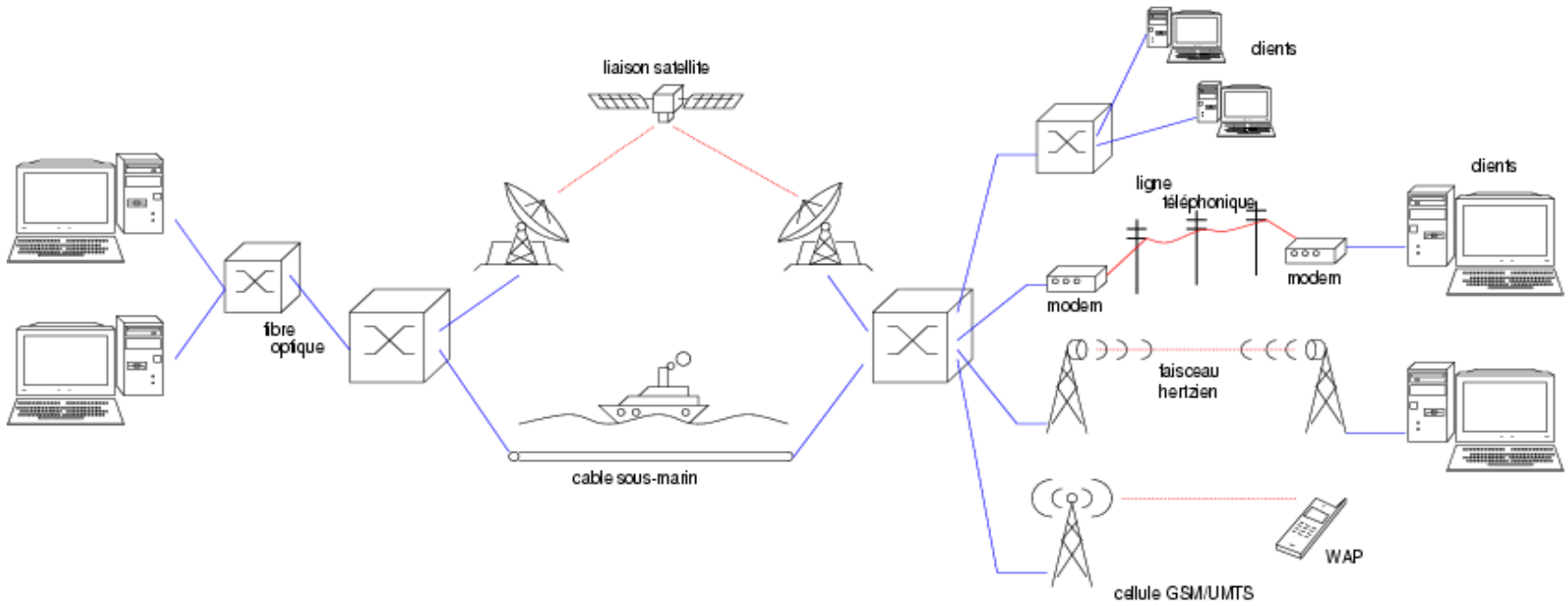
- TCP
  - Fiabilité
  - Reprise sur erreur
  - Beaucoup d’applications : HTTP, FTP…
  - Problèmes : peu de performances, adaptation aux réseaux hétérogènes, TCP friendly…
  - Tout le monde veut le remettre en cause
- UDP
  - Pertes de paquets
  - 5-15% du trafic
  - Peu d’applications…
  - Problème : détruit par les firewalls
  - On voudrait l’utiliser mais difficile

Organismes : nouveaux protocoles

- Un seul organisme gère l'intégration de nouveaux protocoles : IETF
  - Industriels, chercheurs
  - 100 groupes de travail
  - Problème : L'Internet est devenu un réseau opérationnel -> ajouts de nouveaux protocoles difficile

Hétérogénéité

- Performances réseaux
- Equipements réseaux
- Equipements terminaux



Quels protocoles manquent ?

- Dommage :
  - Multicast (presque disponible)
  - QoS
  - Sécurité (IPSec..)
- Tant mieux :
  - Facturation (ça arrive…)

2 points de vues qui s'opposent

- Américains
  - Augmenter la bande passante
  - Redondance, sur-dimensionnement
  - Fibre optique
  - Pas de nouveaux protocoles (ni IPv6, ni QoS…)
- Européens
  - Peu de bande passante donc mieux l'utiliser
  - Besoin de protocoles pour adaptation aux applications

Réseaux actifs

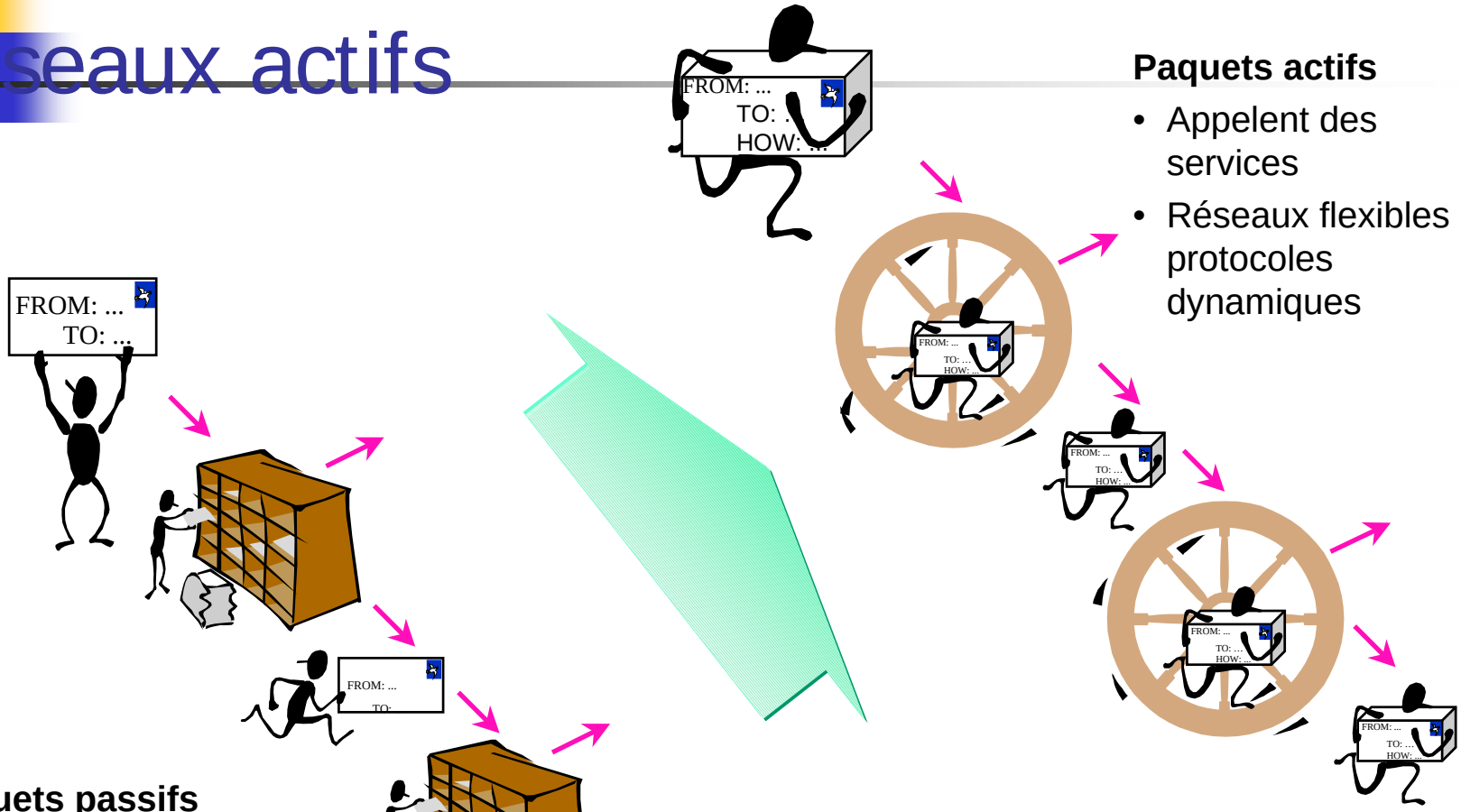
- « Apporter de l'intelligence au cœur des réseaux » pour :
  - Améliorer l'utilisation des ressources réseaux
  - Déployer dynamiquement de nouveaux protocoles
  - Gérer l'hétérogénéité des équipements
  - Prendre en compte l'hétérogénéité des flux de données
  - Proposer de nouveaux services aux applications et opérateurs

Réseaux actifs

Reposent sur :

- La mise en oeuvre de nœuds actifs / routeurs actifs au sein du réseau
- L'intégration d'environnements d'exécution
- Le déploiement de services personnalisés
- Adaptation des applications et des équipements réseaux
- L'inter-opérabilité avec les réseaux passifs

Réseaux actifs



Paquets actifs

- Appelent des services
- Réseaux flexibles et protocoles dynamiques

Paquets passifs

- Pas de services personnalisés
- Réseau rigide et passif

Apporter de l'intelligence dans le réseau pour déployer des protocoles adaptés aux flux de données

=> **Nœuds/routeurs actifs**

De nombreuses approches : Darpa

("out-of-band" code injection)

Discrete Approach → **Configurable Node**

Active Services

CANES

ANN

Active Bridging

ANTS

ANCORS

PLAN

Messenger

Smart Packets

Integrated Approach → **Packet Programming**

("in-band" code injection)

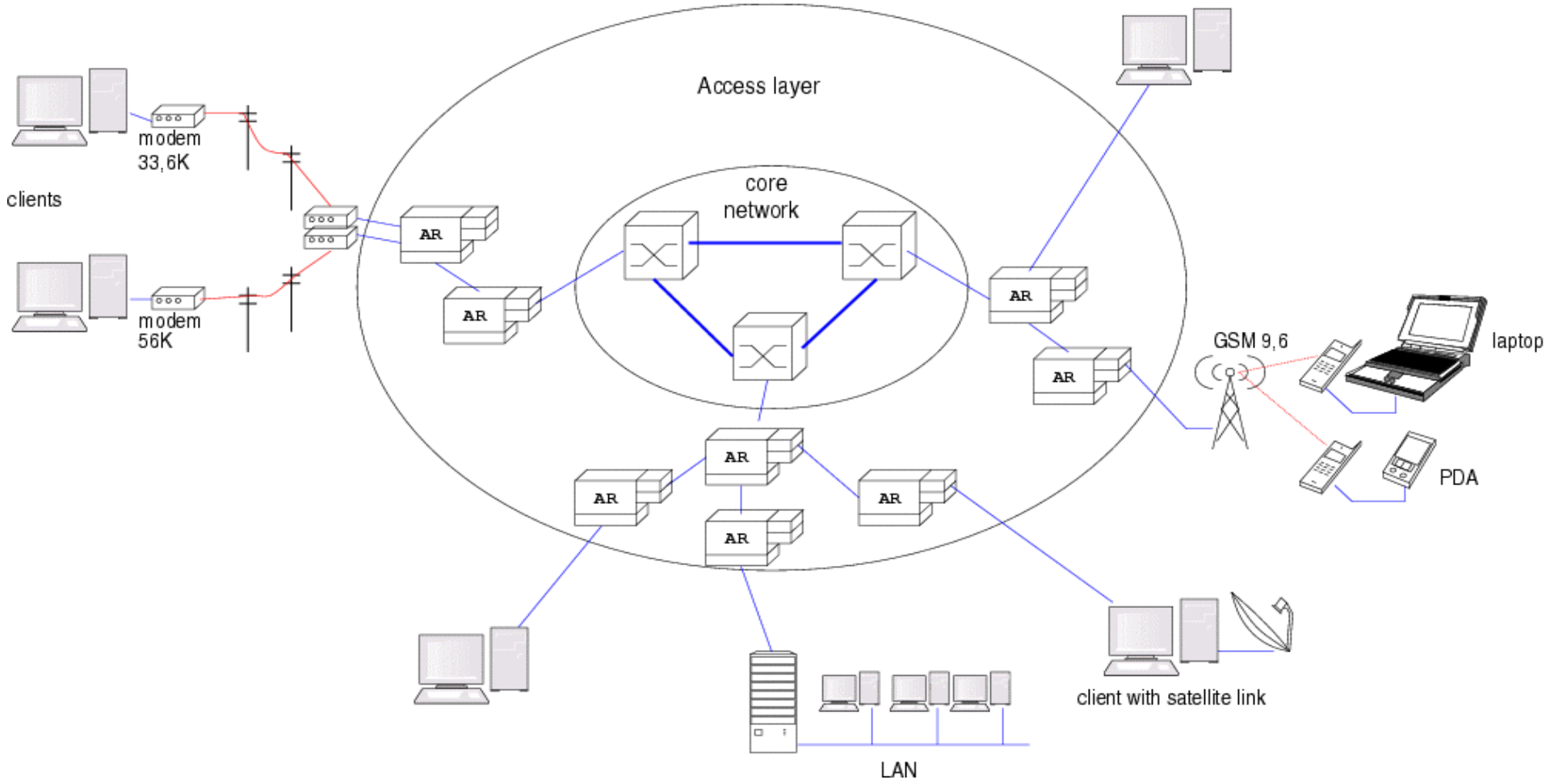
Manques

- Performances
- Sécurité

Routeurs actifs

- Fonctions de routage
- Puissance de traitement (distribuée)
- Software / Hardware
- Rien dans les backbones
- Tout dans les réseaux d’accès

Exemple : Réseaux actifs



Hautes performances et Grille active

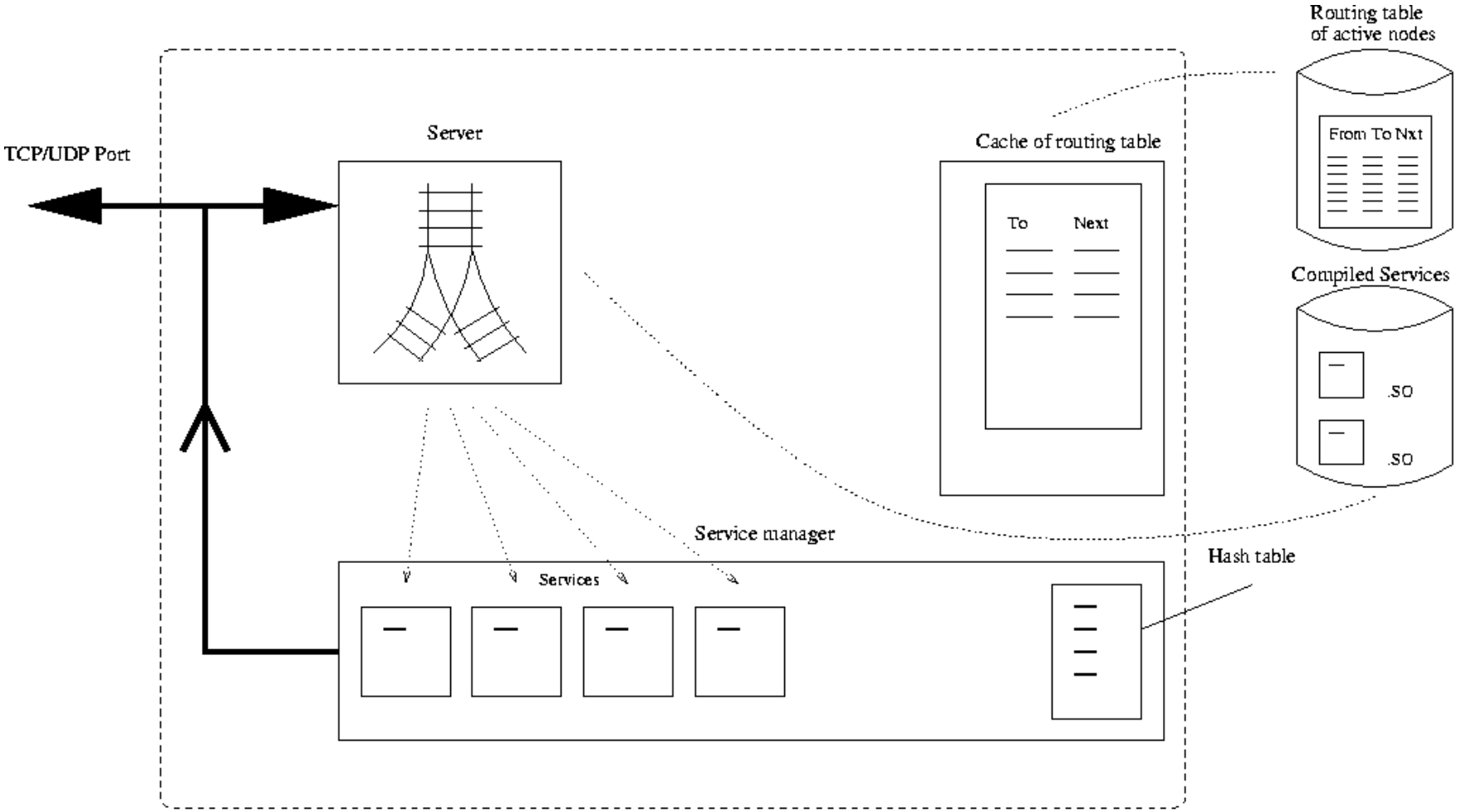
Hautes performances dans les RA

- Performances pour le transport de données
  - Environnements d'éxecution optimisés
  - Compilation / portabilité
  - Hétérogénéité (OS / Réseaux)
- Performances de traitement sur les paquets
  - Multi services
  - Architecture distribuée
  - Traitement à la volée

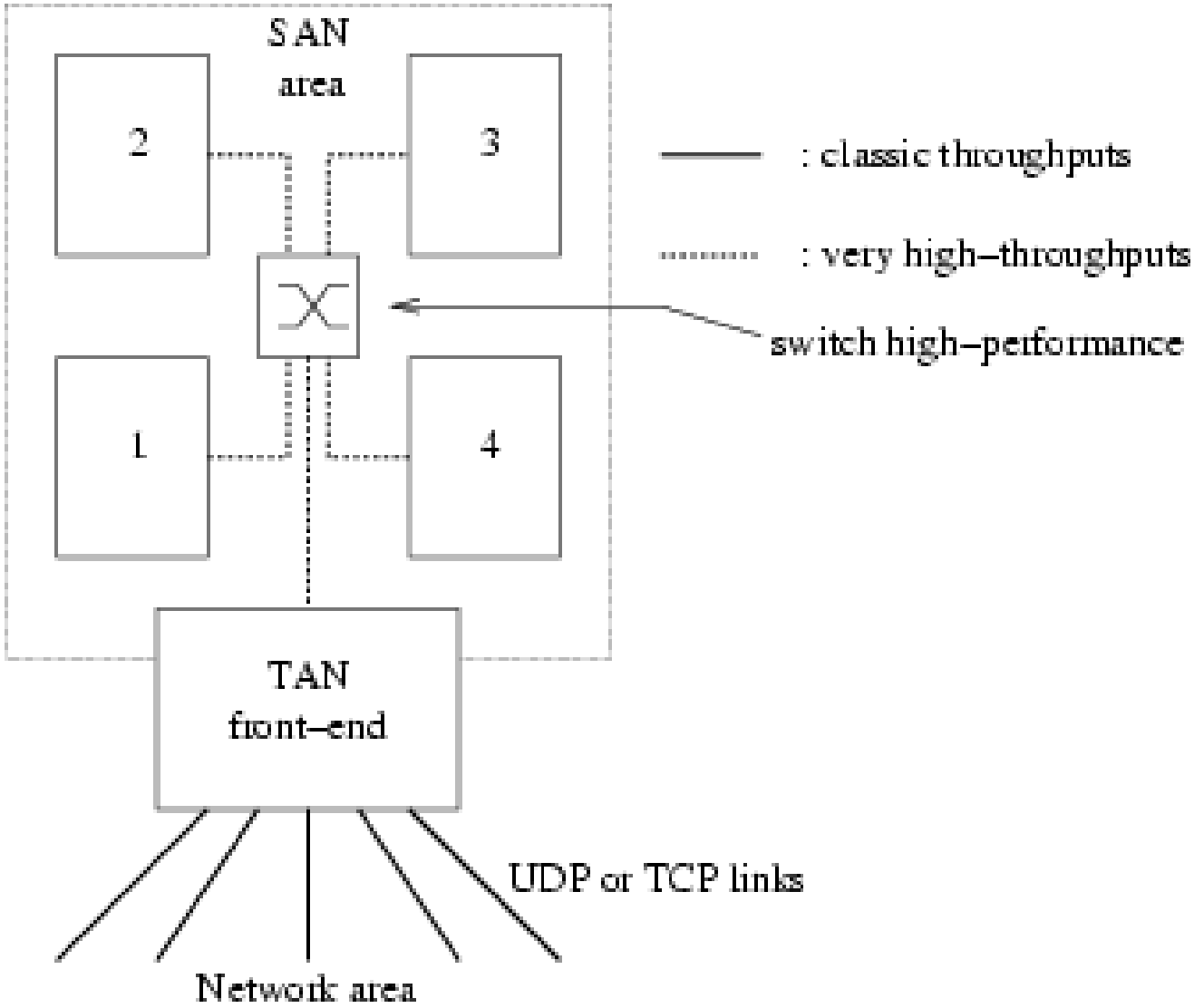
EE actif haute performance: Tamanoir

- Tamanoir : basé sur Java
- Architecture distribuée multi-threads
- Optimisation / compilation (GCJ)
- Paquets actifs : URL de services
- Déploiement dynamique
- Gestion de l'hétérogénéité
- Multi flux : UDP / TCP
- Déport de services au plus proche du fil : Kernel / NIC

Tamanoir architecture

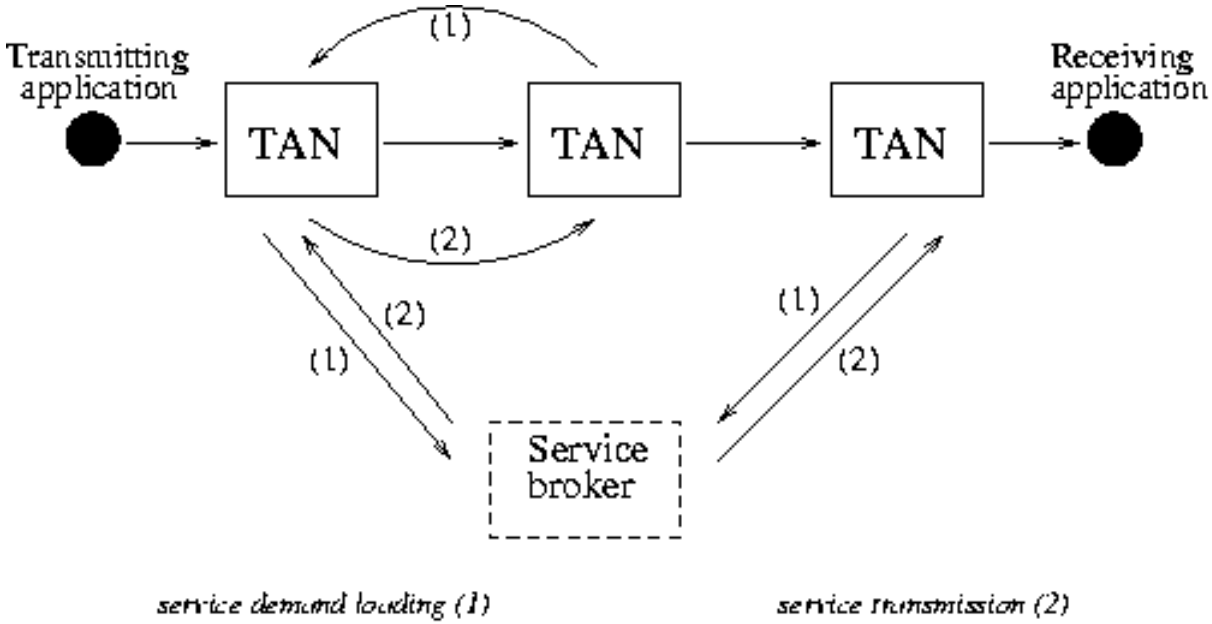


Tamanoir architecture : on a cluster

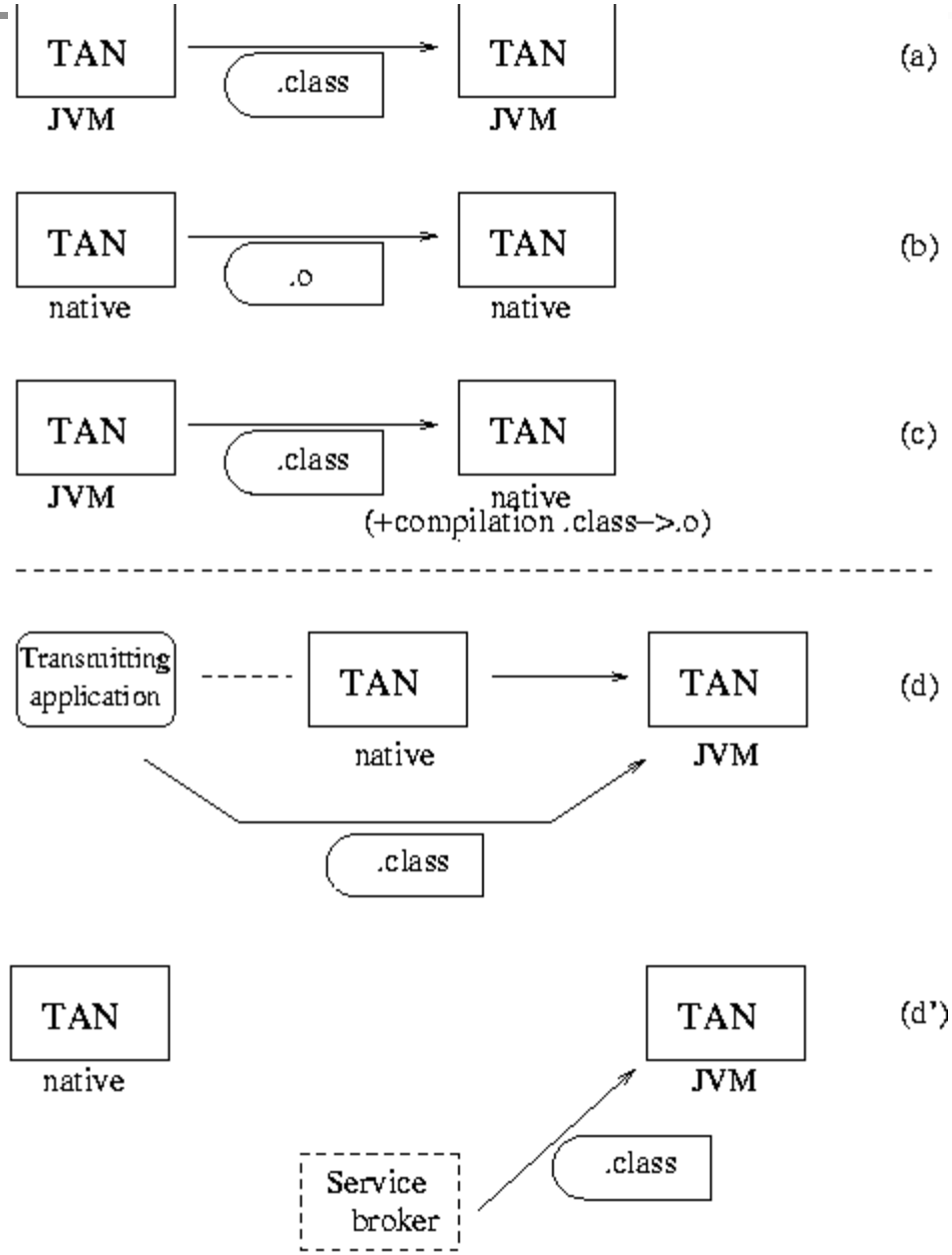


Déploiement de services actifs

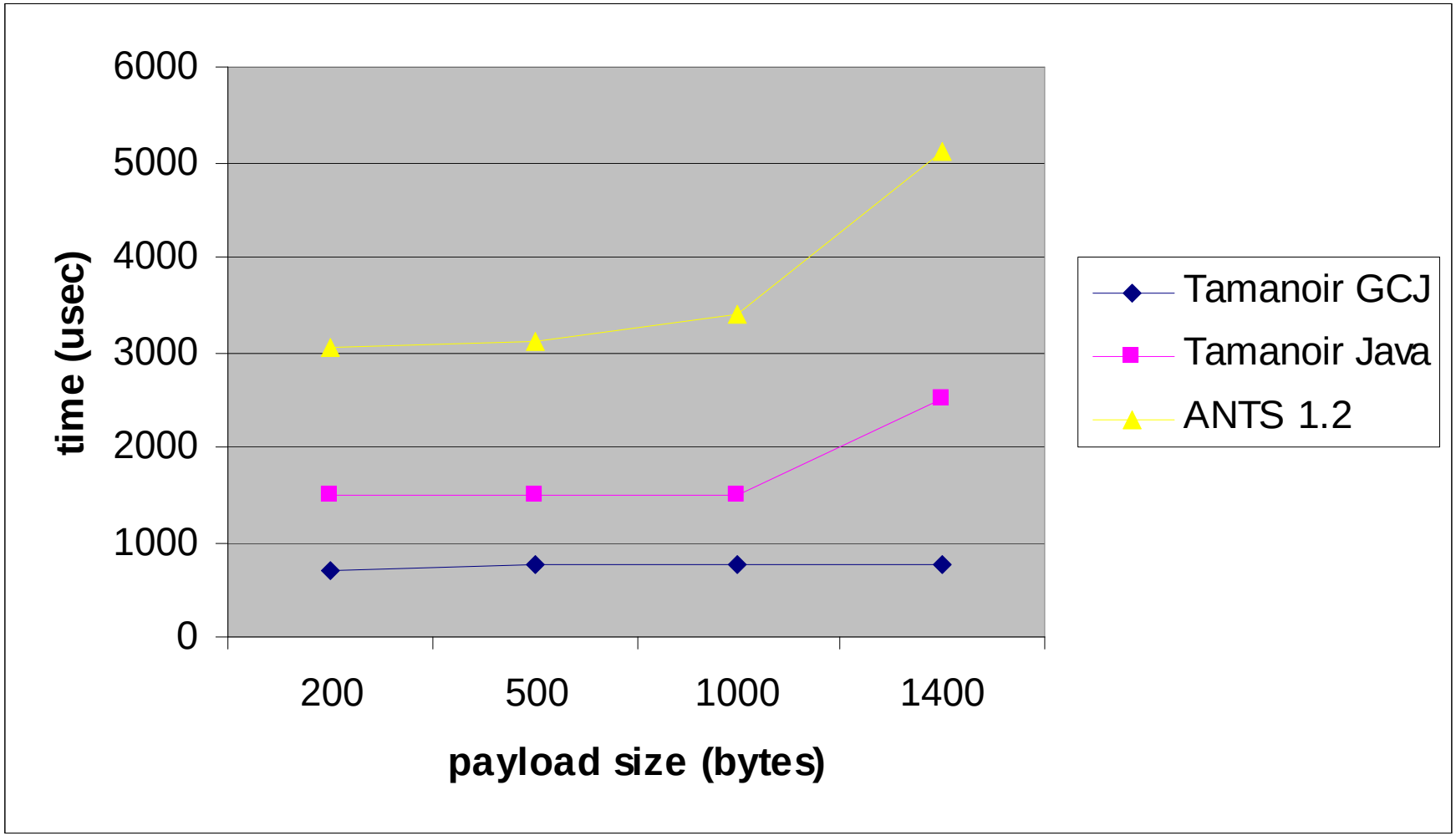
- De l'application / Grid middleware
- Depuis un TAN
- Depuis un Web service Broker



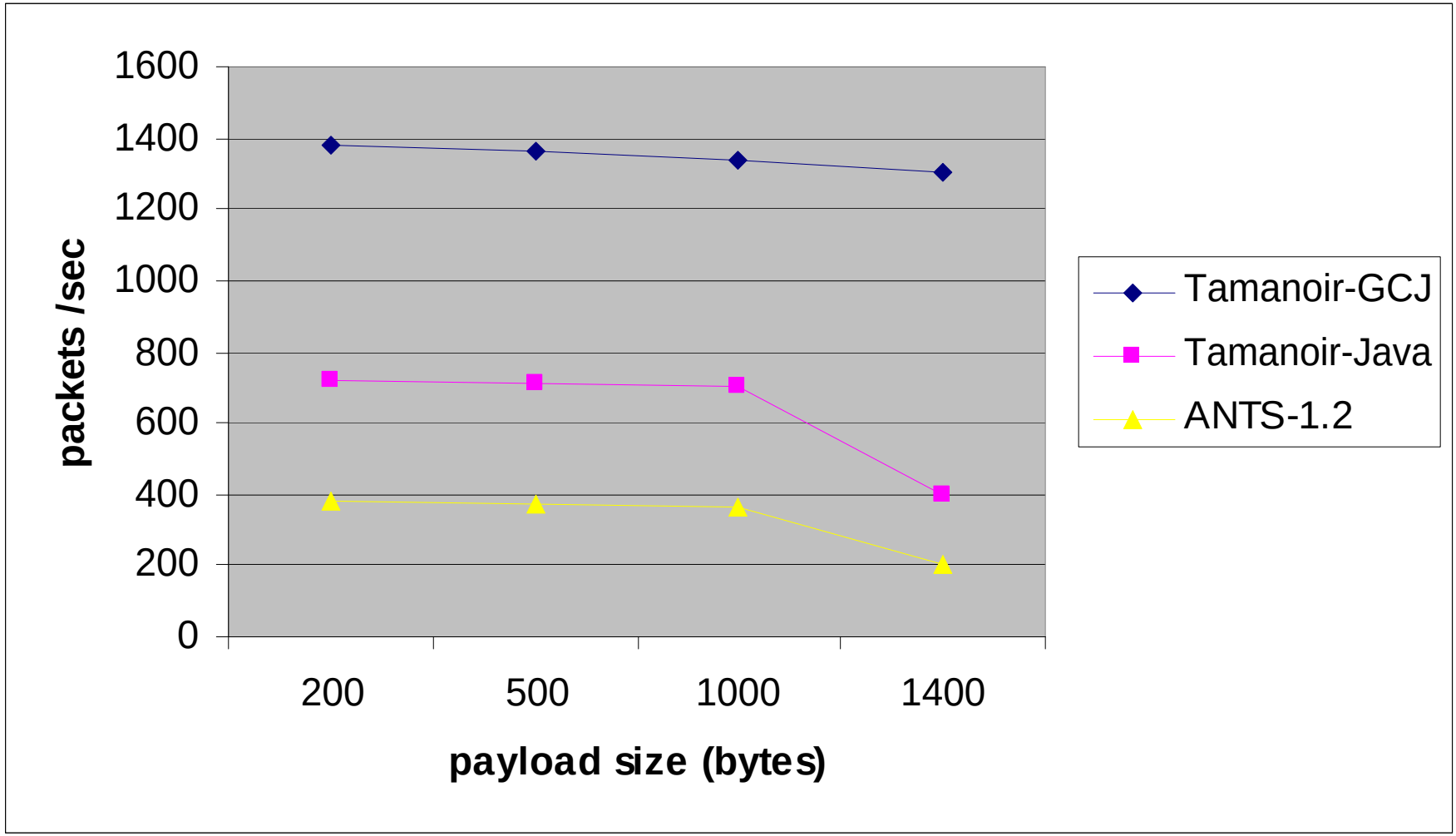
Heterogénéité



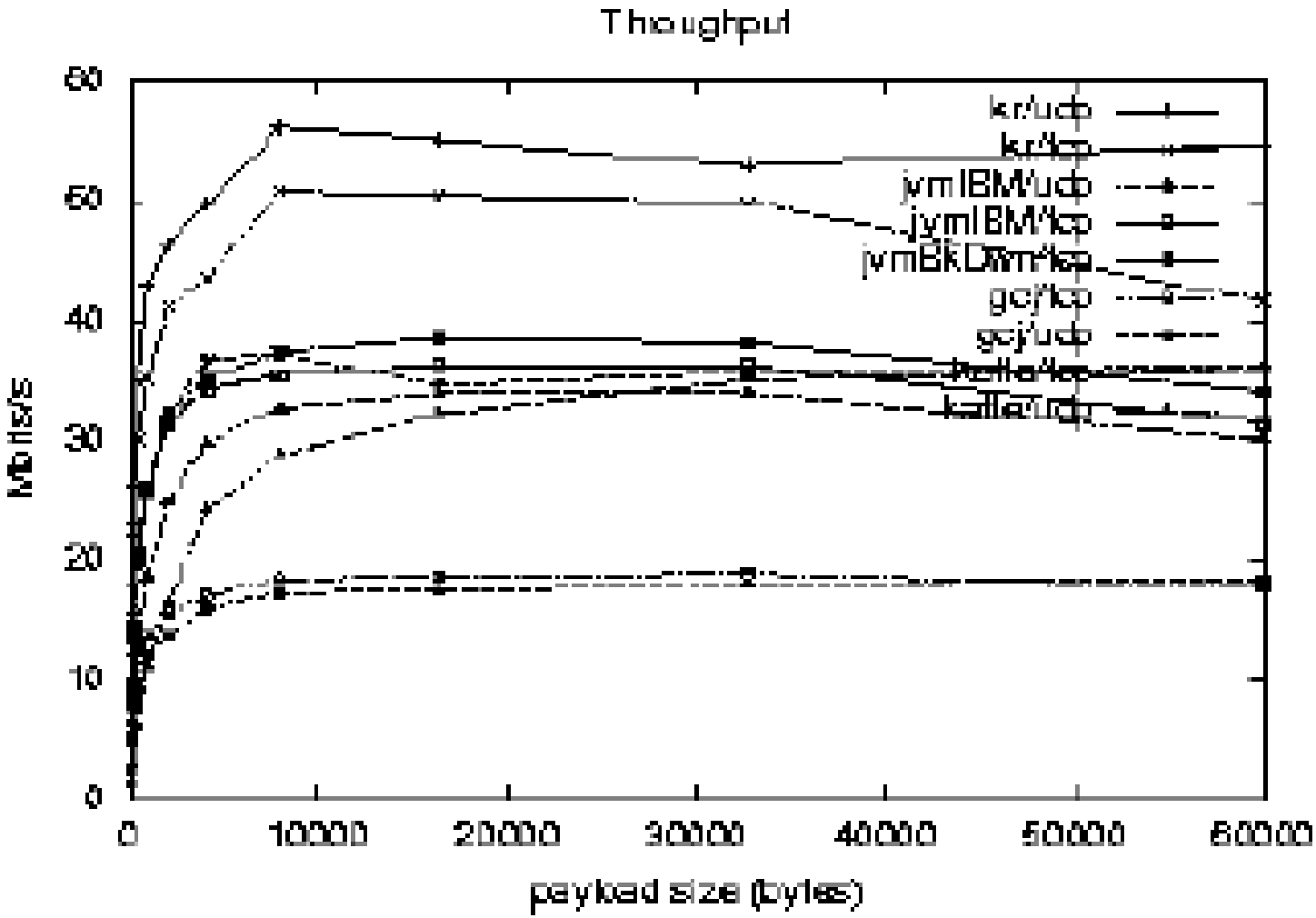
Performances : latency



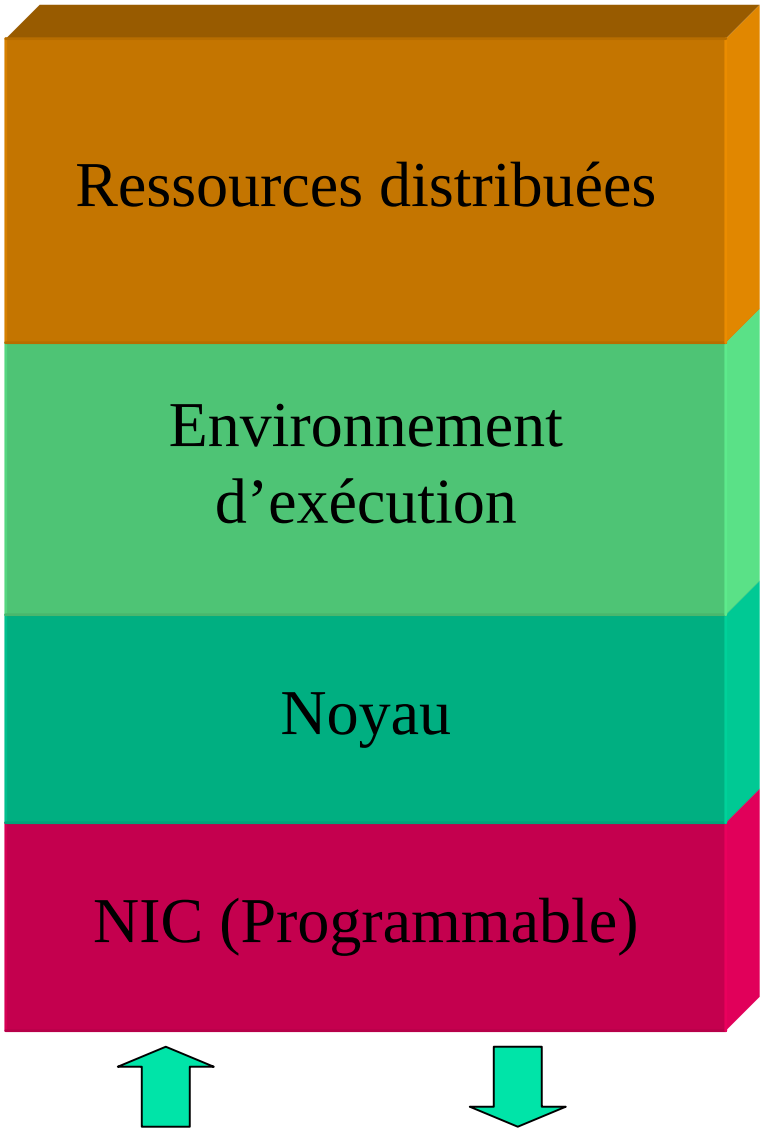
Performances : throughput



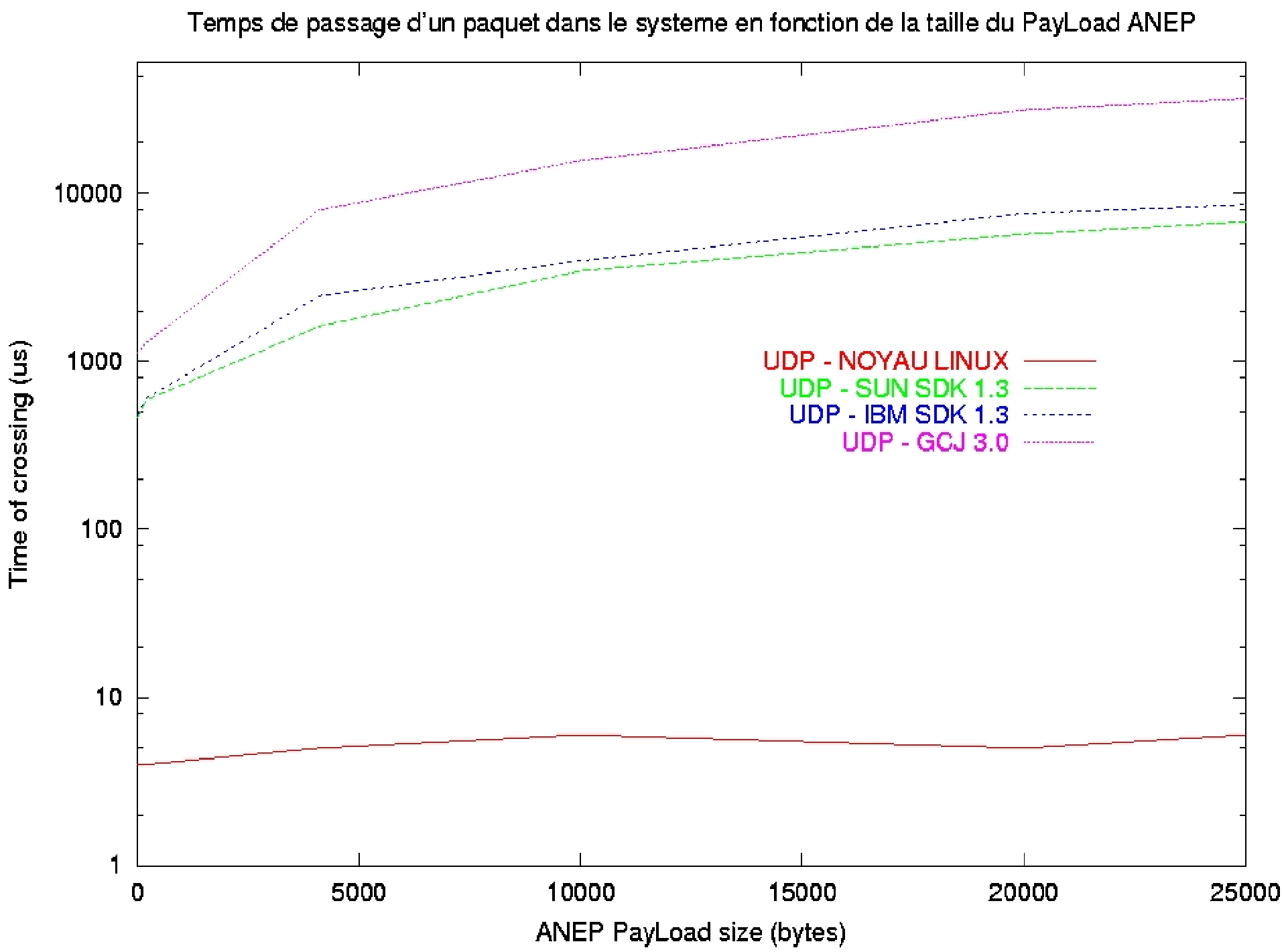
Performances : throughput



Architecture d'un routeur Tamanoir



- Services lourds : stockage distribué, adaptation des flux, compression à la volée….
- Déploiement de services / liens avec Middleware : multicast….
- Services Middle : routage..
- Services réseaux légers (marquage/QoS…)


Temps de passage d'un paquet dans le systeme en fonction de la taille du PayLoad ANEP
Time of crossing (us)
10000
1000
100
10
1
0
5000
10000
15000
20000
25000
ANEP PayLoad size (bytes)
UDP - NOYAU LINUX
UDP - SUN SDK 1.3
UDP - IBM SDK 1.3
UDP - GCJ 3.0

Grille active

GRID computing

- Taches distantes
- Applications distribuées sur réseaux longue distance
- Avec des environnements middlewares (Globus, Legion, Condor, Netsolve, SGE, Nimrod…)

- Protocoles réseaux classiques : IP, TCP

Grid targets

- Multi cluster architecture
  - Set of clusters / parallel machines
  - Few sites
  - Power process sharing

- Mega computing
  - Thousand of machines
  - Cycles stealing

For Grid designers : « networks are simple and do not support Grid applications and middleware »

Grid needs

- Grid control streams
  - Middleware environment deployment
    - Heterogeneity, dynamic topology
    - Dynamic enrollment
  - Application deployment
    - Collective communications : multicast, gather
    - Source deployment, results gathering
    - Fault tolerance
  - Grid support
    - Monitoring, network sensors
    - Node data
    - Enrolled machines information

Grid needs

- Grid applications streams
  - Grid application input / output
    - Parameters
    - data
  - WAN parallel processing
    - Communications between tasks
    - Point2point, global
    - QoS

Grille active

« Un réseau plus intelligent et dynamique pour supporter middlewares et applications de la Grille par des services adaptés. »

Flux de contrôle Grille

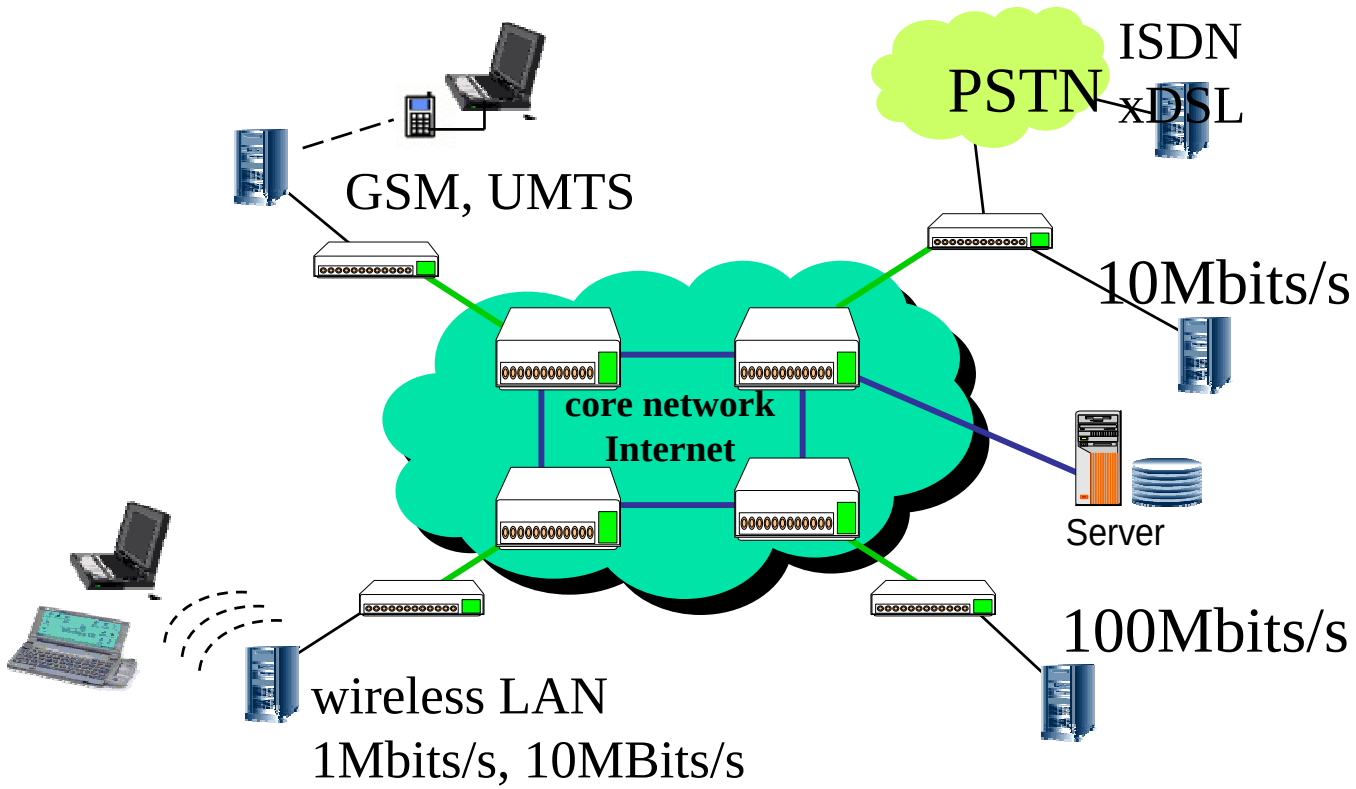
- Déploiement Middleware : hétérogénéité, topologie dynamique, enrôlement…
- Déploiement Applications : communications collectives, diffusion exécutables, filtrage résultats, tolérance aux pannes…
- Support de la Grille : monitoring, informations sur les nœuds…

Flux Applicatifs

- Entrées / sorties distantes : paramètres, données…
- Calcul Grille : communications inter-jobs, QoS...

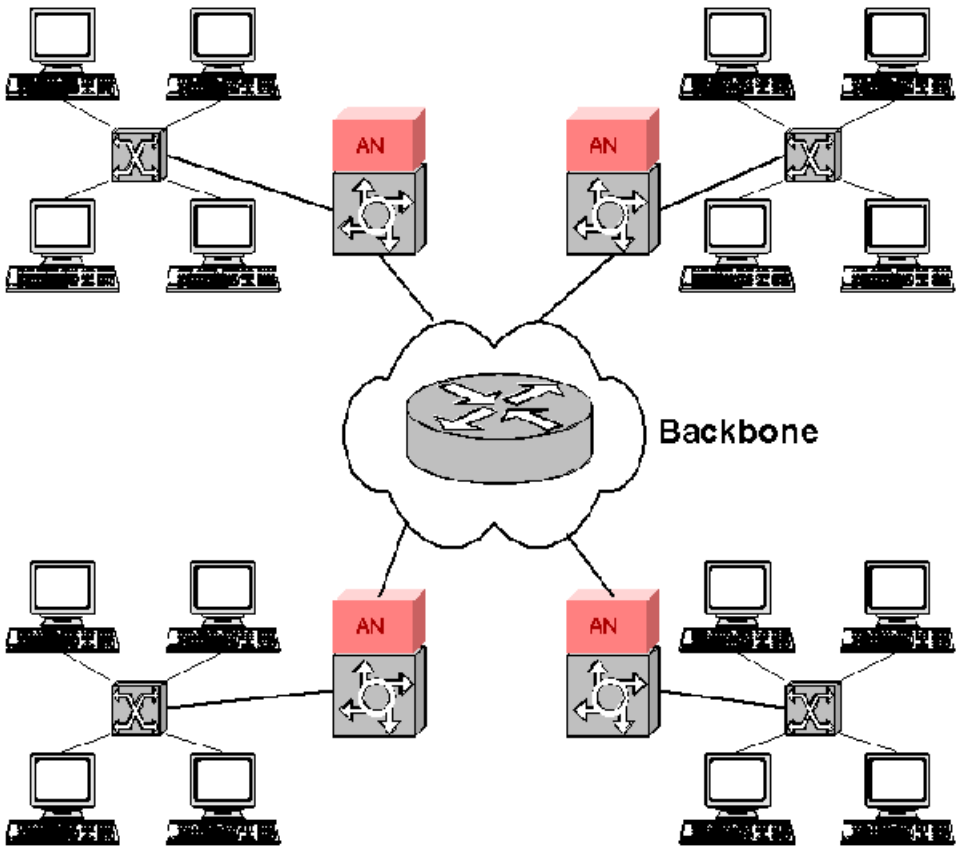
Designing Active Grid on the edge

- Nothing in the backbone (Internet)
- Grid Active nodes on network periphery

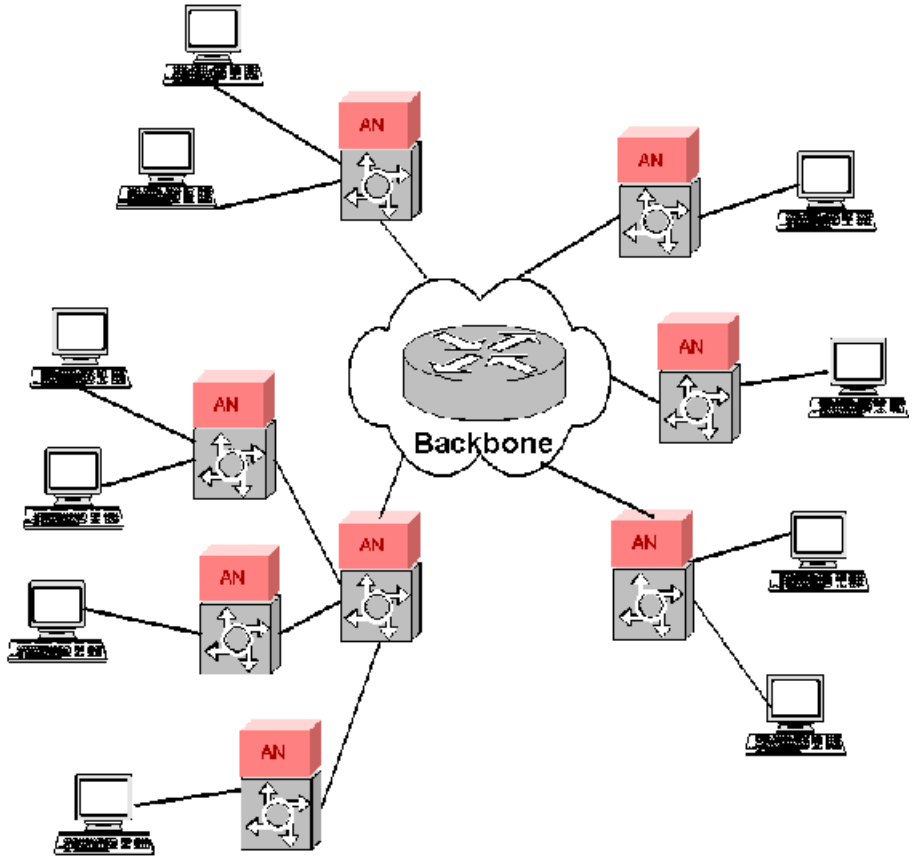


Grille Active

Méta-computing



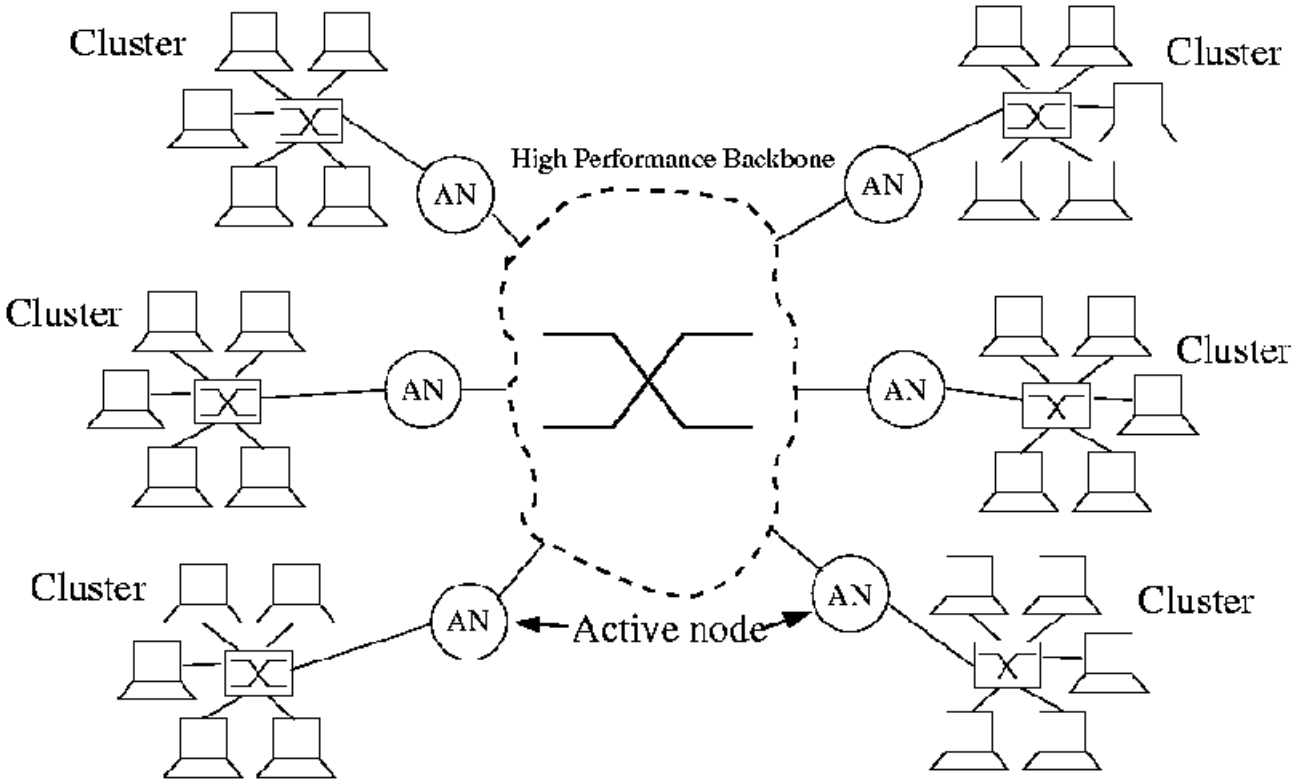
Mega / Global computing



- Rien dans le Backbone
- Nœuds actifs dans les réseaux d'accès

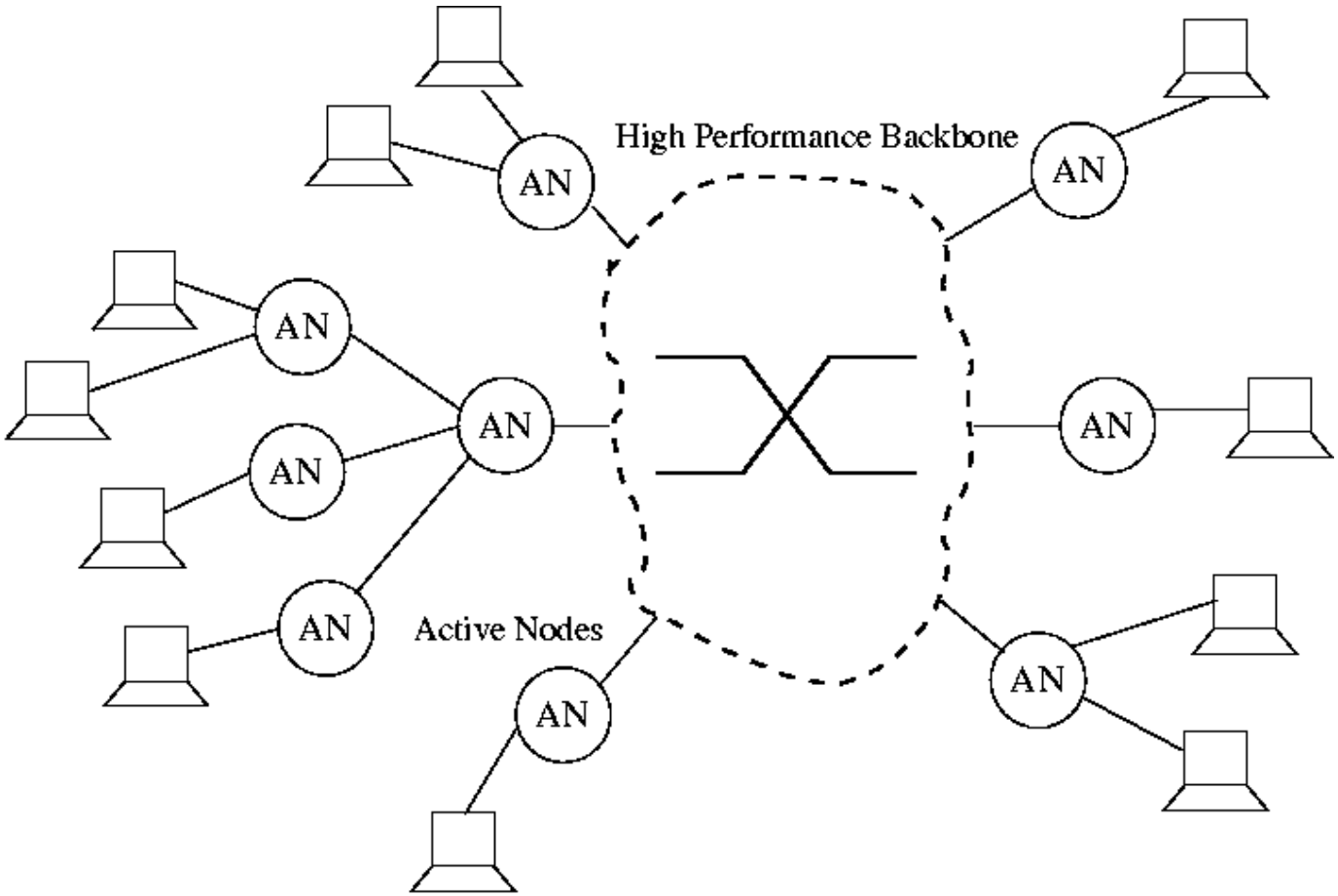
Designing A-Grid : Multi-cluster Grid

- AN : cluster head
- Manage data streams entering and leaving
- Local communications protocols
- Aggregate output streams



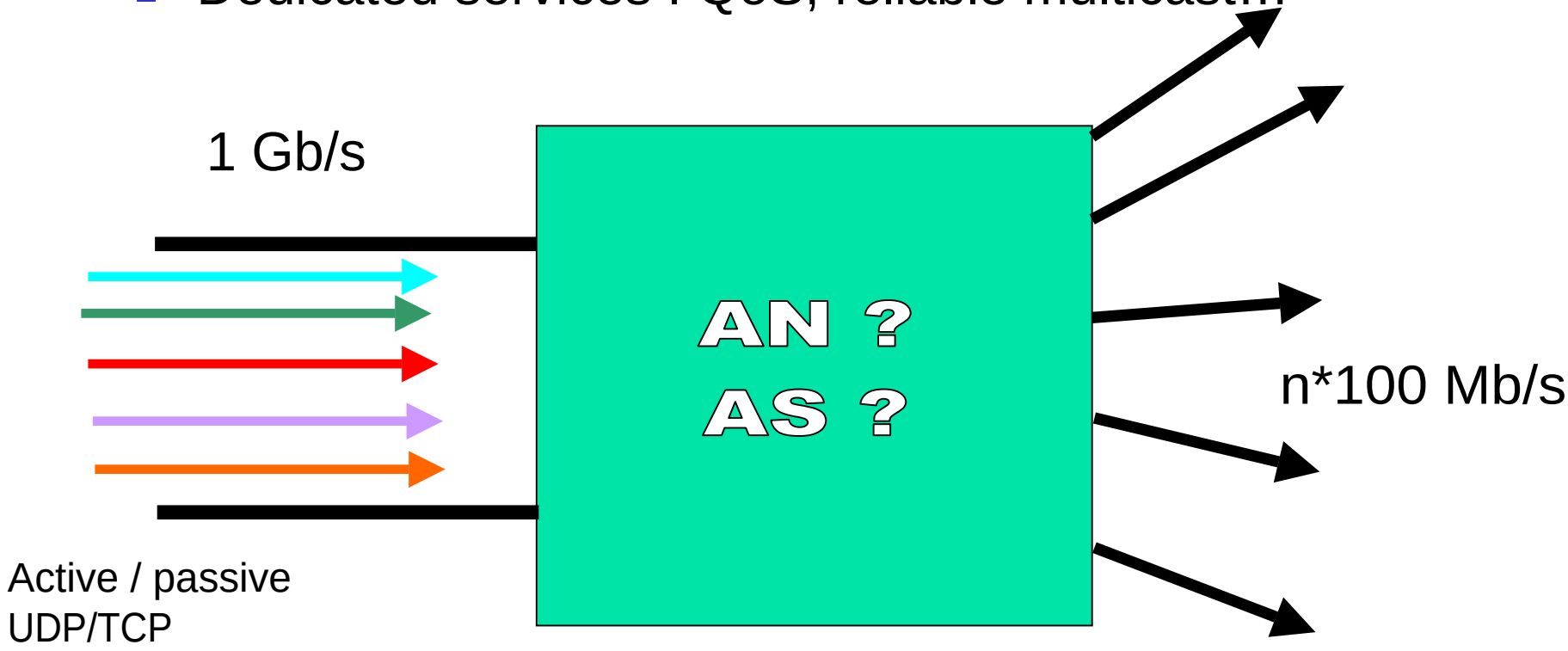
Designing A-Grid : Mega Computing

- AN associated with set of nodes
- AN hierarchies on different networks



What do we want ? AN and AS

- Active node interacting with Grid middleware and applications
- Passive and active packets
- Heterogeneous networks and performance
- Dedicated services : QoS, reliable multicast…



Tamanoir-G

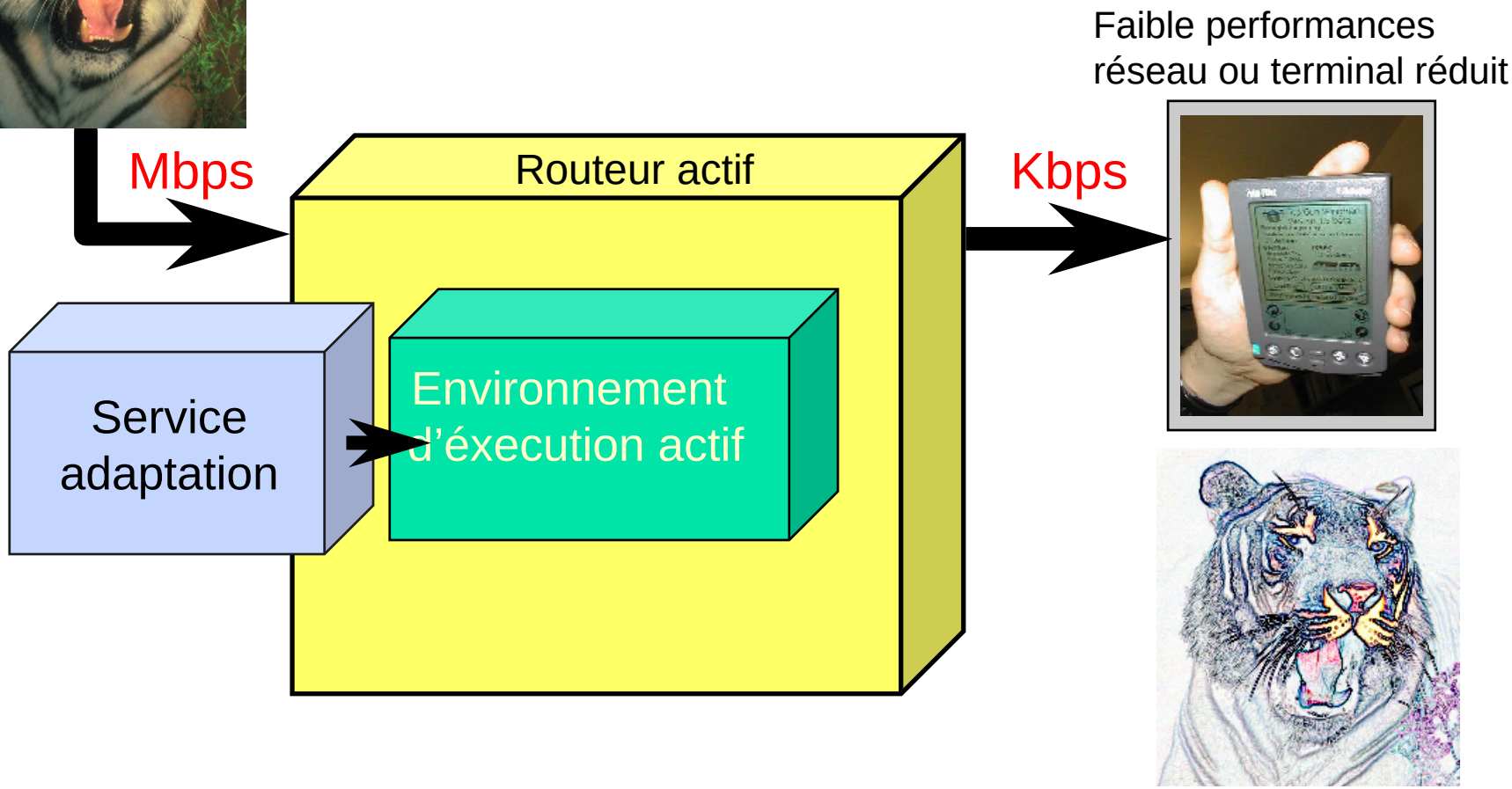
- Gestion multi-flux (UDP, TCP…)
- Interfaçage avec middleware et environnements de Grid computing
- Environnement hautes performances
- Librairies de services Grid (multicast…)
- Expérimentations sur plate-forme locale
- Déploiements de nœuds actifs sur Etoile
- Validation expérimentale avec applications et/ou Middleware

Proposer de nouveaux services : services actifs

- Adapter des flux multimédia
- Sécuriser le transport de données
- Contrer des attaques réseaux
- Cache actif
- Multicast fiable
- Qualité de service

Adaptation de flux multimédia (vidéo)

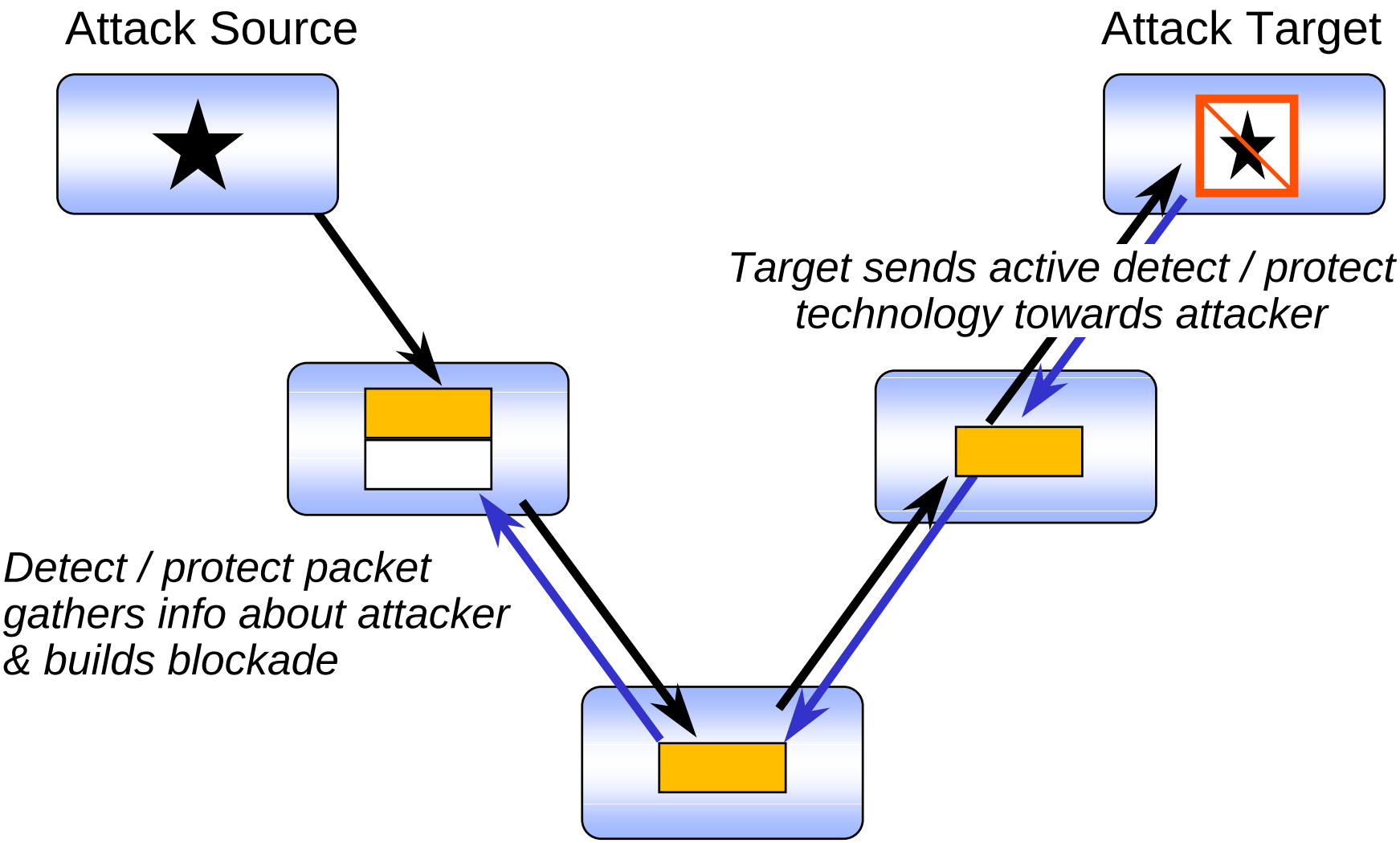
De larges flux transportés sont adaptés dans l'équipement réseau le plus proche



Sécurité déportée dans le réseau

- Fonctions de sécurité trop couteuses
  - Protocoles proposés : IPSec
  - Flux de 100 Mbps sécurisé : 2.5 Ghz calcul
  - Solutions hardware / software
- Déport de la sécurité dans le réseau
  - Grappes
  - Domaines de confiance
  - Passerelles de sécurisation

Contrer les attaques réseaux : DDoS



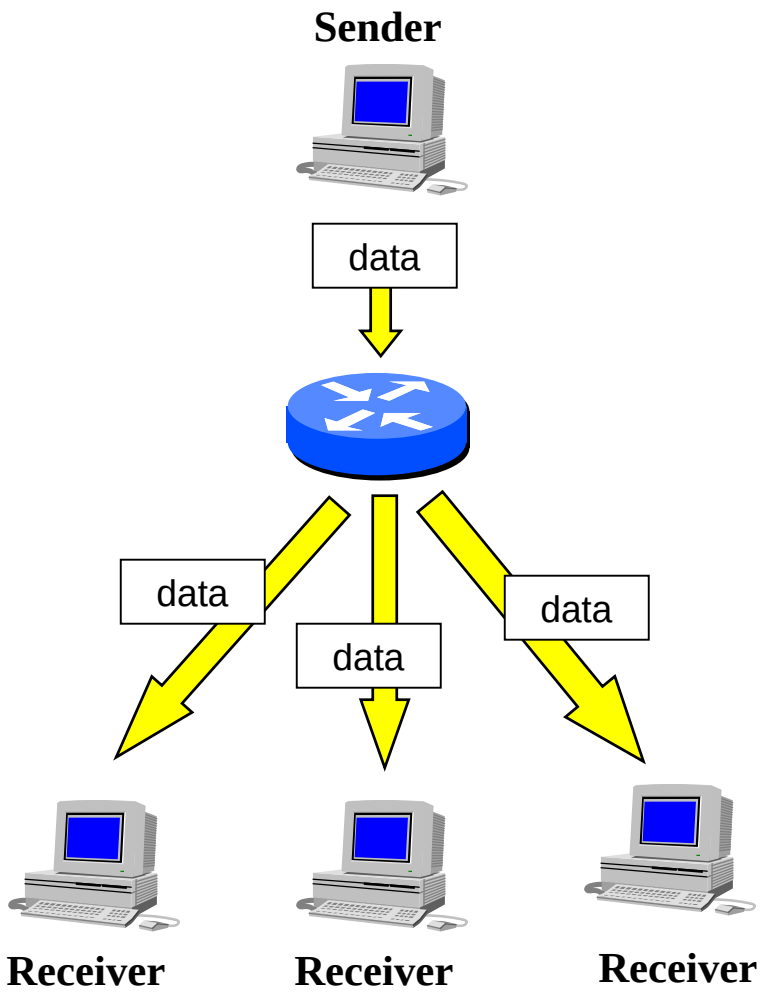
Cache actif

- Proposer du stockage dans le réseau
- Routeurs embarquent des disques
- Stockage logisitique (à la Akamaï / CDN)
- Facile de faire : proxies, cache Web…

Multicast Fiable

Multicast fiable

- Pas n-unicast pour l'émetteur
- Pas de pertes
- Ré-utilisation de IP Multicast
- Distribution de données efficace de 1 vers n
- Distribution de patches, ASP, multi-ftp, push…

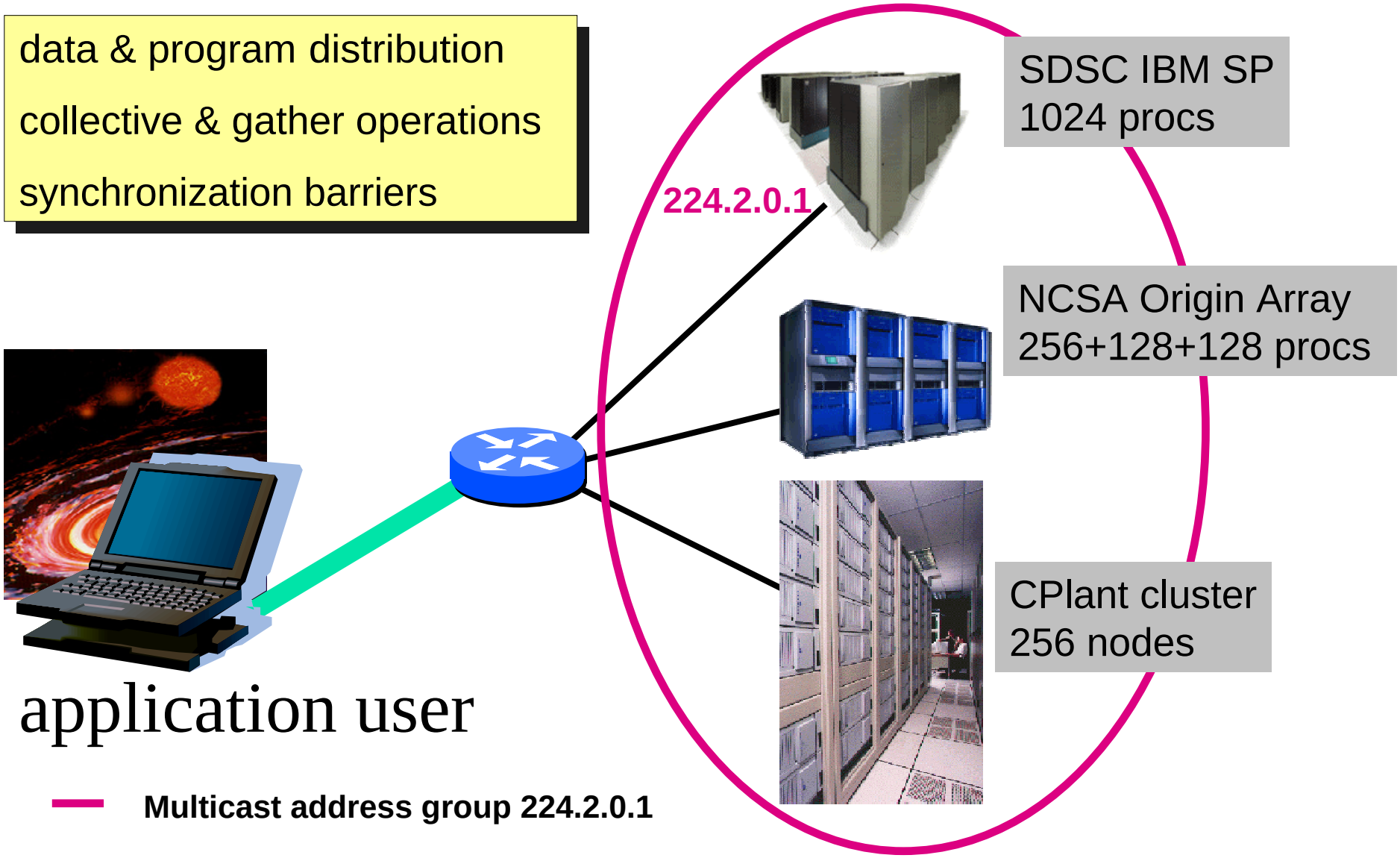


Exemples

data & program distribution

collective & gather operations

synchronization barriers



Multicast fiable

- Problème de la recouverte de perte (loss recovery)
  - feedback (ACK or NACK) implosion
  - replies/repairs duplications
  - Adaptation aux changments dynamiques d'inscriptions
- But
  - Réduire les latences
  - Réduire le trafic de remontée (achnowledges)
  - Diviser les arbres de diffusion pour profiter de la localité

Solutions

- Traditionnelle
  - Retransmissions de bout en bout
  - Suppression des NACK basée sur les receveurs
- Contributions actives
  - cache des données pour retransmission locale
  - aggregation feedback (globale ou locale)
  - subcast

Comparing reliable multicast strategies

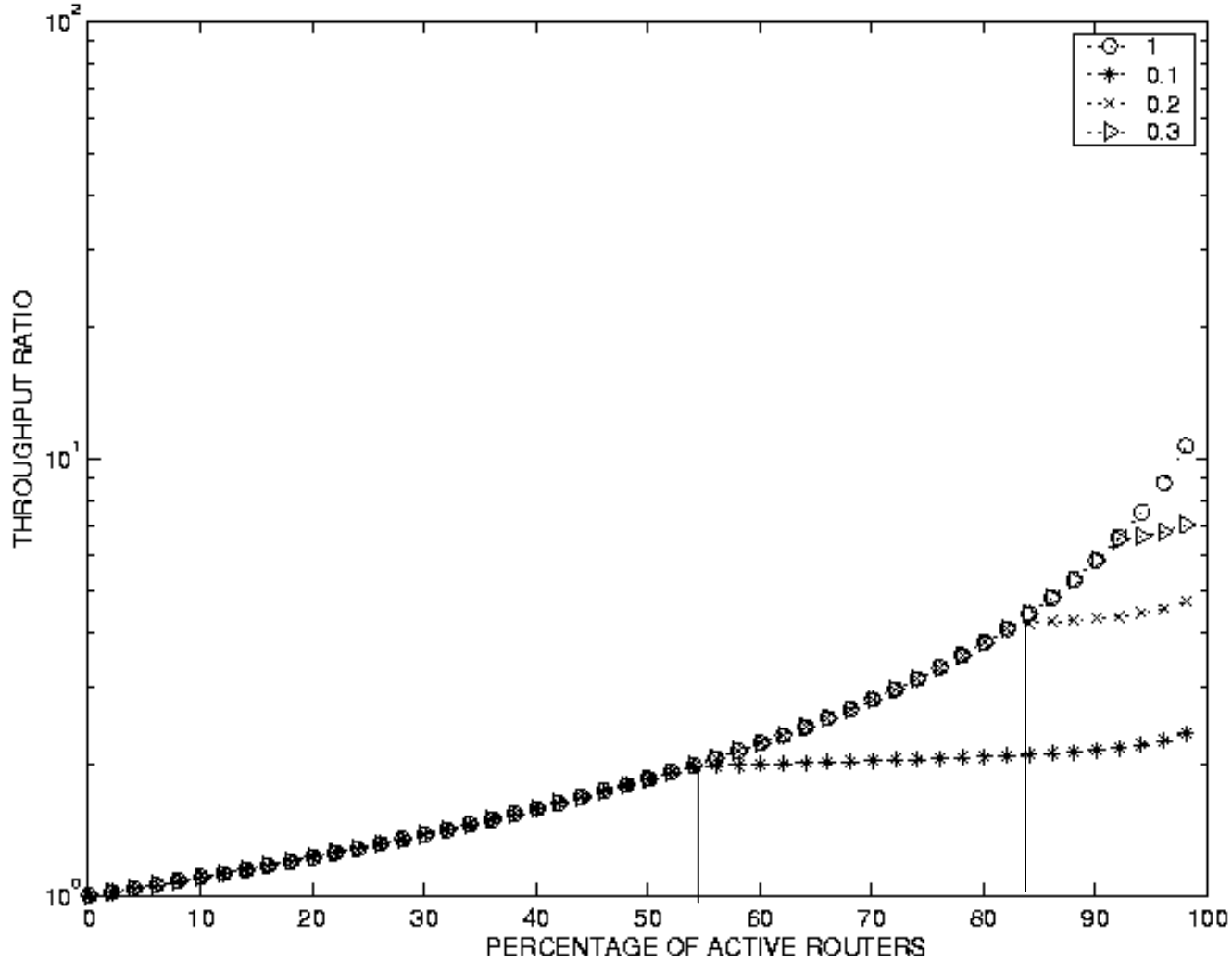
- S1 : global NACK suppression

- S2 : local NACK suppression
- $S2^S$: + subcast from the source

- S3 : global NACK suppression
  subcast from the routers
- $S3^S$: + subcast from the source

Impact of active router density

S3



Active Grid QoS

In-Network QoS or End to End QoS

- **Pure in-network mechanisms**
  - QoS within a single network element (queing, scheduling, shaping)
  - QoS signaling techniques for coordinating QoS from end to end and between network elements
  - QoS policy, management and accounting functions to control and administer end to end traffic across a network.
- End to End mechanisms
  - To mask the deficiencies of the network
  - QoS adaptation: retransmission
  - Examples: TCP, adaptive applications

In-Network QoS Approaches

- Two types of QoS approaches available:

  A) ***Resource reservation*** *:* network resources are reserved according to an application's QoS request, and subject to bandwidth management policy. Requires a Reservation protocol

  - Explicit signalization (ex. RSVP protocol)

  B) ***Prioritization*** *:* network traffic is classified and network resources are shared according to bandwidth management policy criteria. To enable QoS, classifications give preferential treatment to applications identified as having more demanding requirements.

  - Implicit signalization (packet marking)

QoS service levels

- QoS services differs in how tightly they can be bound by specific bandwidth, delay, jitter and loss characteristics.
  - Best effort: lack of QoS
  - Guaranteed Service (**hard QoS**): an absolute reservation of network resources for specific traffic

    => **IntServ Architecture**
  - Differentiated Service (**soft QoS**) : some traffic is treated better than the rest.

    => **DiffServ Architecture**

Grid and QOS

- Will Grid applications benefit from the currently proposed in-network QoS services?
- GRID QoS needs are:
  - delivery of a complete data file is guaranteed, the priority of each individual packet is not.
  - The user specifies the ultimate delivery time when the data transfer must complete.
  - Predictable throughput on demand
  - Efficient management of heterogeneous streams

Grid Networking Issues

- backbones will provide lot of bandwidth (10Gb/s in the core for GEANT in the DATAGRID project for ex.)
  - but which added value services ? (MPLS?)
  - lot of inter-domain issues (LANs)
- Application/middleware requirements are heterogeneous
  - high energy physics, earth observation, biology have different needs…
  - streams are difficult to identify and to model.
- Monitoring Information is important
  - Integrate measurements into Grid information service.
  - To provide information for Middleware

Active QoS

Chosen approach :

- Upload the pure in network mechanism in the active nodes (eg. Intserv or DiffServ)
  - Active DiffServ Model.
  - The architecture is based on a DiffServ or over provisioned backbone
  - Active QoS adaptation is made at the edge.

Active GRID QoS

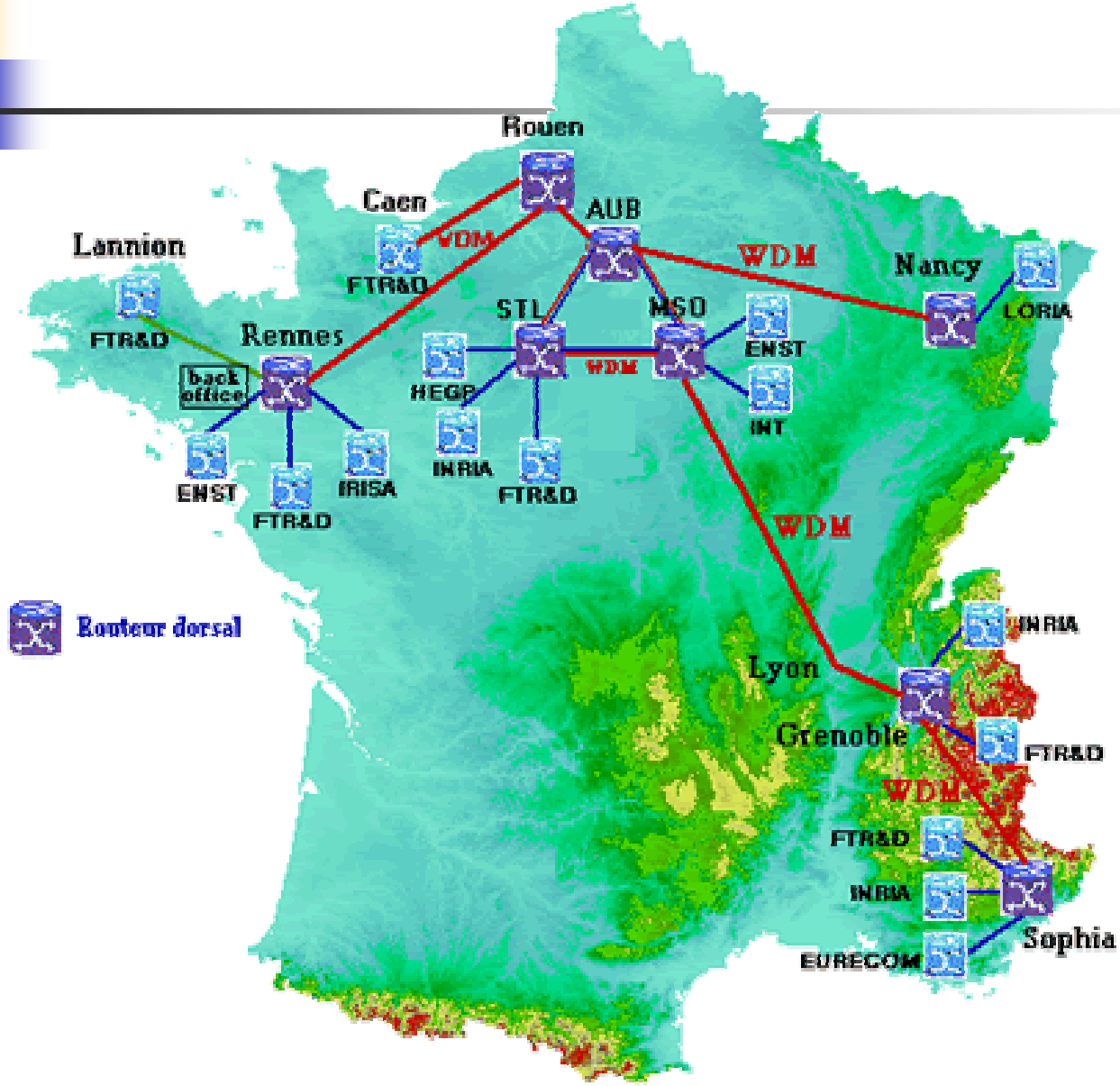
- In our Active DiffServ model, the first active node is responsible of the packets marking.
- An ADSCP (*Active DiffServ Code Point*) is inserted in the ANEP capsule, just before the application payload. The length of this field is variable.
- Following nodes process packets using ADSCP, general status of the network, local status of the node, in order to optimize the PHB (per hop behavior).
- Processing can be
  - Specific (vs systematic) dropping
  - Filtering (shaping, discarding)
  - Adaptative TOS field remarking depending on the network/flow status when entering a DiffServ domain
- Requires distributed and active monitoring and marking services at the edge.

Quelques projets en réseaux actifs

- RNRT VTHD++
- RNTL EToile

RNRT VTHD+ +

- Projet VTHD : Réseau Vraiment Très Haut Débit
  - IP/WDM
  - SONET / SDH : 2.5 Gbits. Ethernet 1 Gibt/s en sortie
  - Internet « nouvelle génération »
  - acteurs : France télécom, INRIA, ENST, INT, Institut Eurécom
  - Expérimentations de la technologie active haute performance
  - http://www.vthd.org

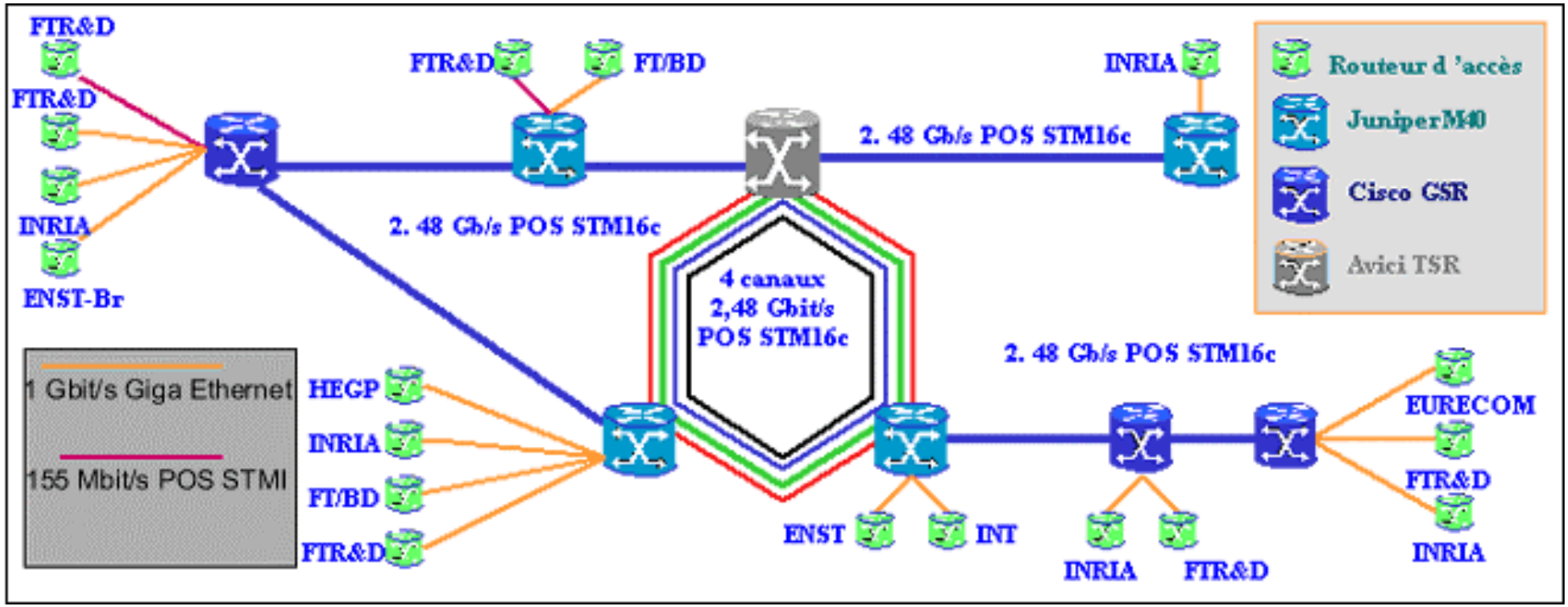

Rouen
Caen
AUB
Lannion
FTR&D
WDM
WDM
Nancy
LORIA
Rennes
STL
MSD
ENST
back office
HEGP
WDM
INT
ENST
FTR&D
IRISA
INRIA
FTR&D
WDM
Routeur dorsal
INRIA
Lyon
Grenoble
FTR&D
WDM
FTR&D
INRIA
Sophia
EURECOM

Des systèmes de transmission optique multi-longueur d'onde en cœur de réseau

Le réseau VTHD a adopté pour sa partie dorsale cette architecture simplifiée IP/WDM où les routeurs sont directement interconnectés par les systèmes de transmission optiques WDM. En associant sur une même fibre optique, plusieurs porteuses optiques ou longueur d'onde, la capacité de transmission des fibres déployées dans les réseaux actuels peut être portée à une fraction de térabit/s (10^{12}bit/s). De plus, les systèmes de transmission WDM permettent de gérer avec souplesse la capacité de transmission du réseau de transport par mise en exploitation progressive de longueurs d'onde supplémentaires.

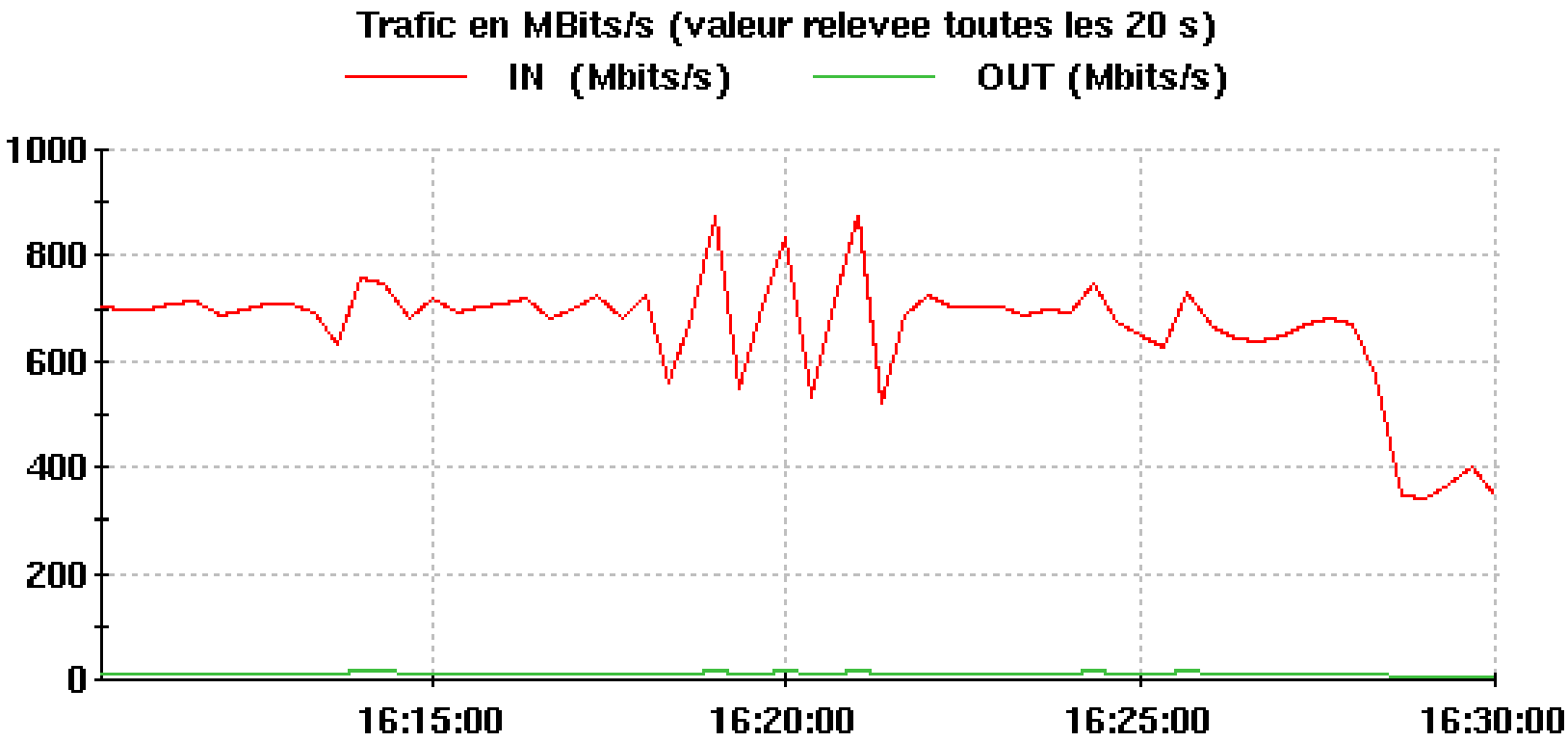
Les longueurs d'onde interconnectant les routeurs du réseau dorsal VTHD portent chacune un débit de 2,48 Gigabit/s. Un tramage numérique SDH de ces longueurs d'onde permet d'exploiter les porteuses optiques pour y transmettre des flux numériques binaires.

Actuellement, quatre longueurs d'onde sont exploitées sur l'anneau métropolitain parisien portant ainsi le débit d'interconnexion des 3 routeurs de cet anneau à 10 Gbit/s. Les autres routeurs sont interconnectés à 2,5 Gigabit/s.


FTR&D
FTR&D
INRIA
ENST-Br
FTR&D
FT/BD
INRIA
2. 48 Gb/s POS STM16c
2. 48 Gb/s POS STM16c
Routeur d 'accès
JuniperM40
Cisco GSR
Avici TSR
4 canaux
2,48 Gbit/s
POS STM16c
2. 48 Gb/s POS STM16c
1 Gbit/s Giga Ethernet
155 Mbit/s POS STMI
HEGP
INRIA
FT/BD
FTR&D
ENST
INT
INRIA
FTR&D
EURECOM
FTR&D
INRIA

Circuit - Evolution
March, 8 2001

Liaison : Sophia->Grenoble



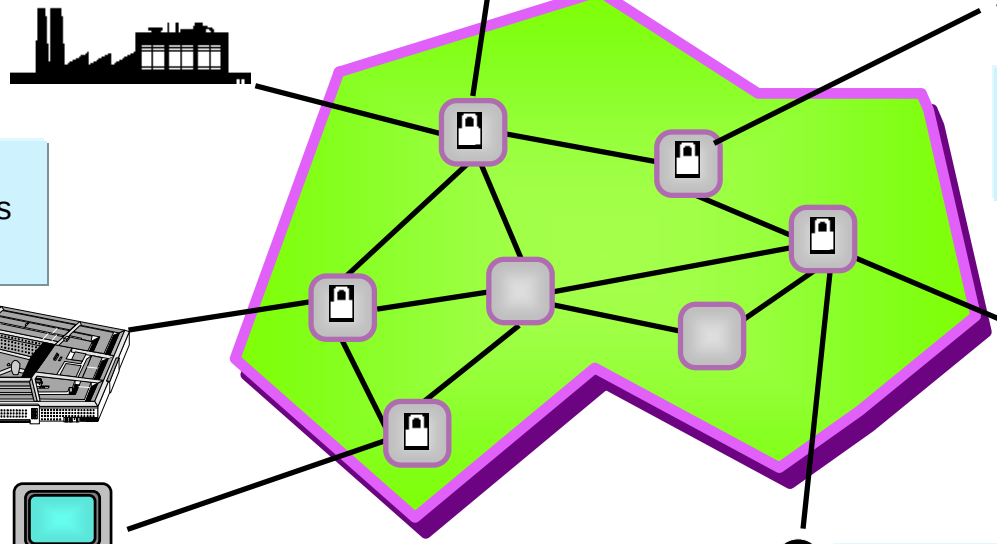
RNTL Etoile

- Proposer, implémenter, valider le concept de Grille Active
- Partenaires : INRIA, UREC, IRISA, SUN, EDF, CS, CEA…
- Middlewares
- Services
- Applications

WHAT IS ALL THIS GOOD FOR?

Injection of new services and technologies is faster and easier

Shrink-wrapped router software



Bandwidth/computation trade-off to support wireless users

"Time to Standards" reduced from ~10 years to months

Fusion/merging of information within the network

Adaptive monitoring, management, and control of network resources

Network infrastructure that can instantly be tailored to mission requirements

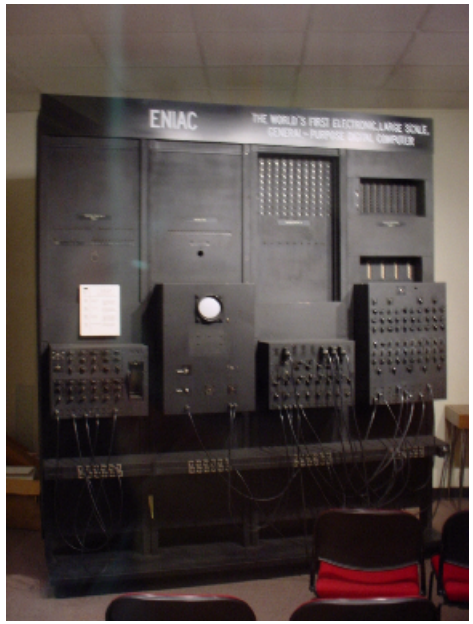
Conclusions

- A-Grid can improve Grid applications :
  - Application deployment
    - ARM : source code, binaries
    - Active cache : gathering, loss recovery
    - QoS : data transfer
  - Grid support
    - A-Grid provides information to middleware
    - Distant tasks management
  - WAN process
    - Active QoS for streams
    - Efficient transport

Conclusion – travaux en cours

- Proposer un routeur actif hautes performances pour un déploiement industriel
- Les réseaux actifs sont à la recherche de la *killer application* : Grid Computing bon candidat
- En cours :
  - Beaucoup de travaux à mener avant apparition grand public (4-5 ans)
  - Intérêts industriels (routeurs actifs)
- More information :

http://www.ens-lyon.fr/~llefevre/Tamanoir



ENIAC, Upenn 2/10/01

This document was created with Win2PDF available at http://www.daneprairie.com.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.