Improving Load Balancing with Multipath Routing

Pascal Mérindol

Jean-Jacques Pansiot, Stéphane Cateloin

ICCCN 2008

Table of Contents

- Multipath Routing Paradigm - Related Works
  - Goals and general context
  - Source routing
  - Hop by Hop routing
- Dijkstra Transverse at depth p
  - Dijkstra Transverse computation DT
  - Validation process DT(p)
- Load Balancing

  Related Works & Traffic Engineering module
- Evaluation
  - Path diversity
  - TE results

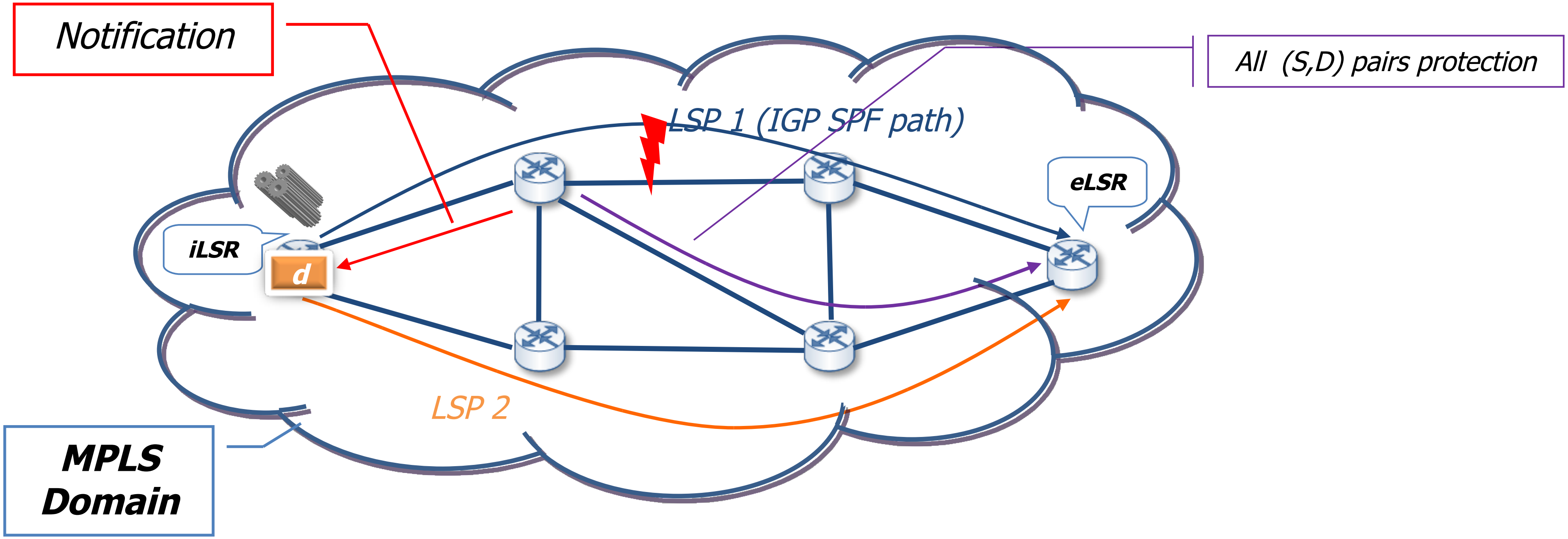
Multipath Routing

Outlines

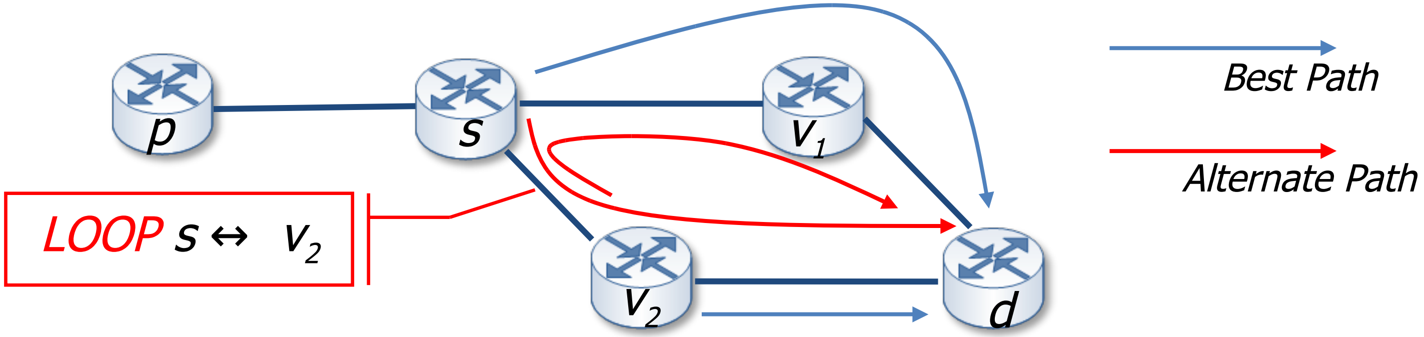
- Purposes and objectives
  - *Load balancing – circumvent congestions*
  - *Protection and restoration - circumvent failures*
  - ➢ Increase throughput and reduces delays
  - ➔ Two ways : **source** or **hop by hop** routing
- General scheme
  1. *Path computation & signalization or validation (loop free paths)*
  2. *Path or link (global or local) traffic analyse*
  3. *Load balancing policy*
  4. *Traffic splitting*

MPLS based scheme
Pseudo-Source routing

- Multi protocol label switching *(**MPLS**)*
  *Explicit path signalling mechanisms such as **RSVP-TE** or **CR-LDP***
- Additional label Switch Paths *(**LSP**)*
  *Since the ingress Label Switch Router (**LSR**) towards the egress LSR*



Multipath hop by hop routing
Related works



LOOP FREE PROPERTY

Loops can be condidered at two levels :

- <u>At node level</u>
  - *Equal Cost Multipath Routing : ECMP*
  - *Downstream Criteria (One hop vision) : OMP-OSPF, LFI , etc.*
- <u>At link level</u>
  - *Two Hop vision depending to the incoming interface*

Validation rules

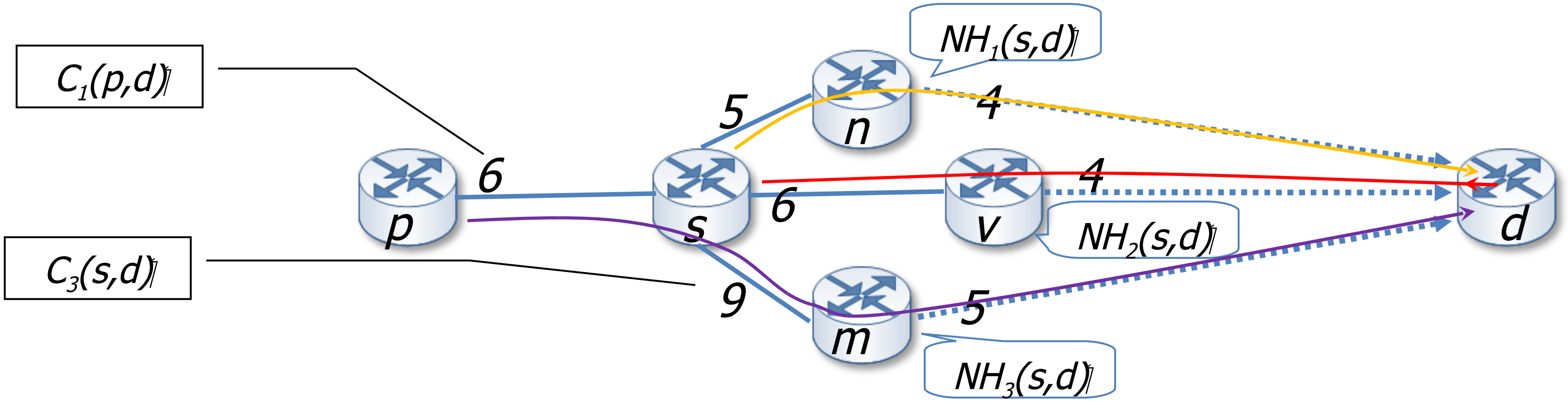
Loop free conditions

An alternate next hop v is viable if :

ECMP : $C_j(s,d) = C_1(s,d)$ / v= $NH_j(s,d)$

Downstream Criteria (LFI) : $C_1(v,d) < C_1(s,d)$

Two Hop vision : $C_1(v,d) < C_1(p,d)$



$C_j(s,d)$: j^{th} best cost computed on s towards d

$NH_j(s,d)$: j^{th} next hop computed on s towards d

Dijkstra Transverse

Path computation

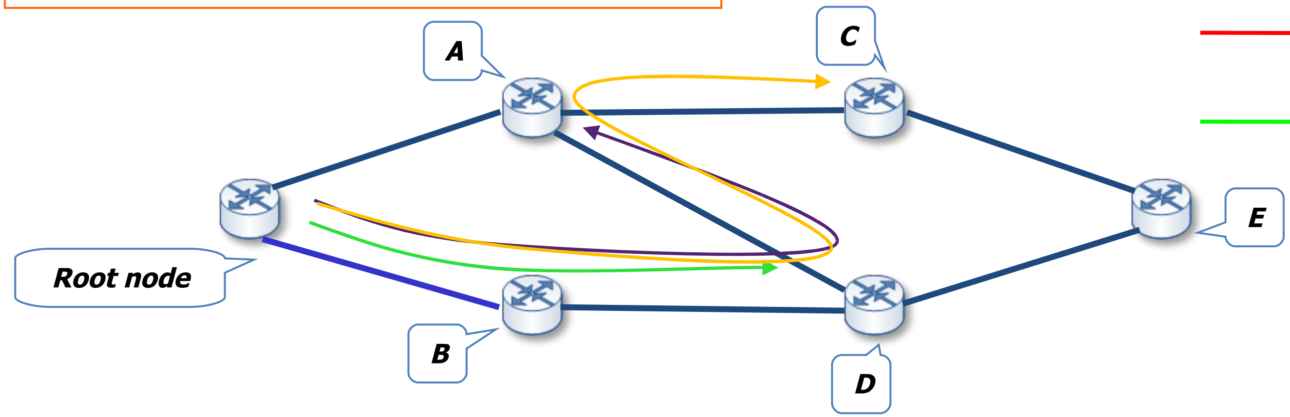
- Dijkstra Transverse (DT) is a enhanced SPT algorithm
- DT computes at least one alternate next hop to every destination

DT computes four sets of paths

- *Best cost path*
- *Simple transverse path*
- *Backward transverse path*
- *Forward transverse path*

Next Hop candidates are store in a cost matrix

- *2 dimensions :*
- cardinal of the successor set of the root
- cardinal of the destination set



Shortest path tree
(with a lexicographical order)

Transverse edge

	A	B	C	D	E
A	1	3	2	2	3
B	3	1	4	2	3

DT(p) Validation process

Routing row validation at depth 1

- Validation at the granularity of the **incoming interface**
- The incoming interface loop free criteria

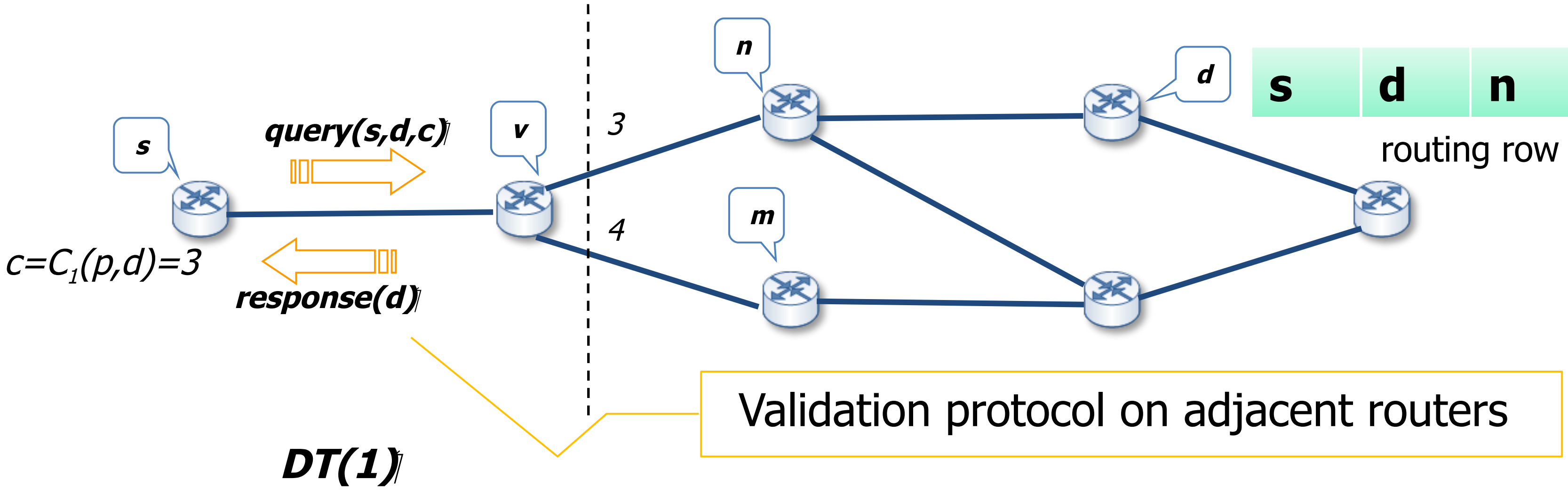
$$C_j\,(v,d) \;\leq C_1(s,d)$$

candidates on v

d	n	3
d	m	4

s	d	n	3

routing row



Validation protocol on adjacent routers

DT(p) Validation process

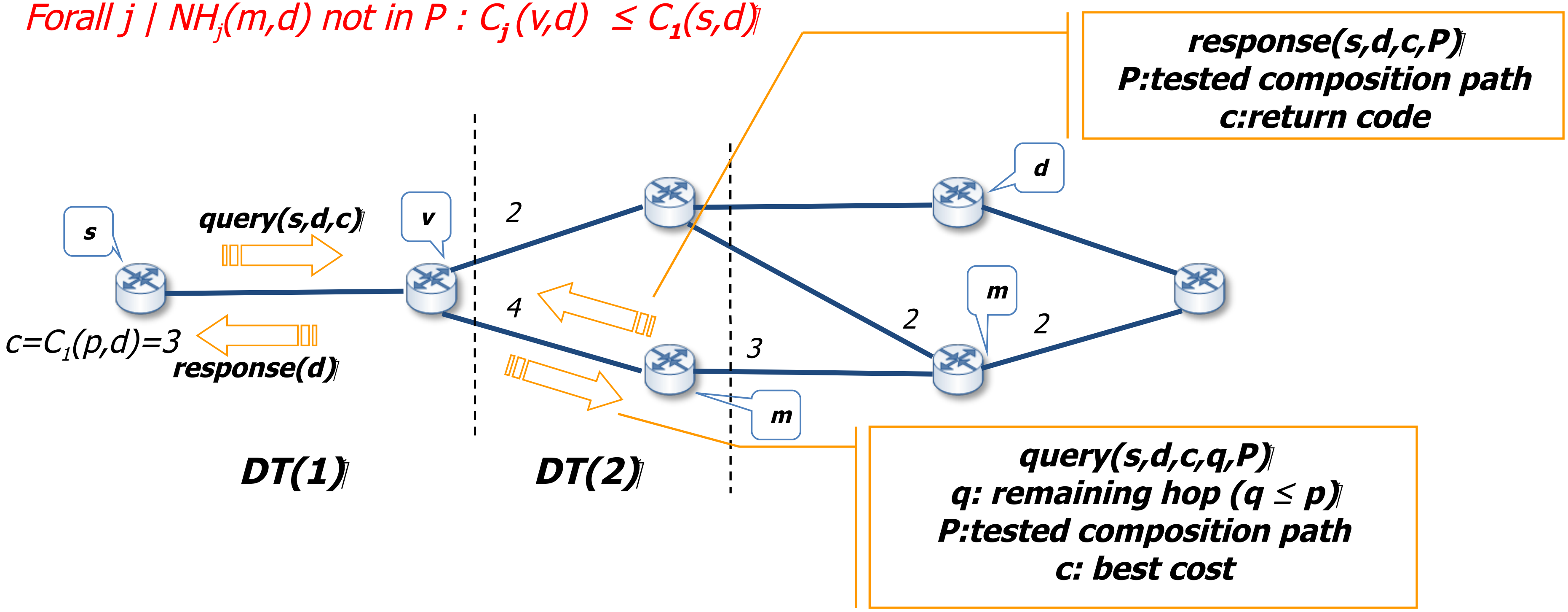
Routing row validation at depth p

- Validation process extension

Each path composition P is evaluated with a wave of request (On the exemple, p=2 and P={v})

- The loop free criteria becomes

Forall j | $NH_j(m,d)$ not in P : $C_j(v,d) \leq C_1(s,d)$



response(s,d,c,P)
P:tested composition path
c:return code

query(s,d,c,q,P)
q: remaining hop ($q \leq p$)
P:tested composition path
c: best cost

Load balancing

Related works and TCP incidence

- Load balancing
  - Modifying links weights *– [Fortz & Thorup, Wang & al., Sridharan & al., etc]*
  - Optimization statement with logical paths *(e.g MPLS) :*

    OFFLINE *(with traffic matrix) oblivious or/and normal case routing – [Applegate & Cohen, Zhang & al., COPE, etc]*

    ONLINE *(with probe protocols) – [MATE, TeXCP, etc]*
  - Incremental heuristics for hop by hop routing *– [ECMP, Vutkury & Garcia-Luna-Aceves, OSPF-OMP, Gojmerac &al., etc]*
- Traffic splitting (without disordering TCP packets?)
  - packet level : *round robin, probabilistic, etc*
  - flow level : *<src,dst,port...> hash function, tag, hybrid, etc.*
  - burst level *- [FLARE]*

Load balancing
Our TE module – basic proportional routing

- Constraints : proportions integrity

For each destination d,
a proportions vector : $\left\{x_1^d(p),\ldots,x_j^d(p),\ldots,x_n^d(p)\right\}$

$$\forall p \sum x_j^d(p) = 1$$

- Objective : minimize the maximum link utilization

Utilization ratio of an outgoing link l on a router s :

$$U(l) = \sum_{\forall d\in N, \forall p\in pred(s)} \frac{x_j^d(p)\times V_d(p)}{c_l}$$

traffic coming from p to d

If $x_j^d(p)$ corresponds to outgoing link l

$$\min \ \max U(l)$$

- Local load shifting incremental process

Two parameters for up and down reaction thresold :

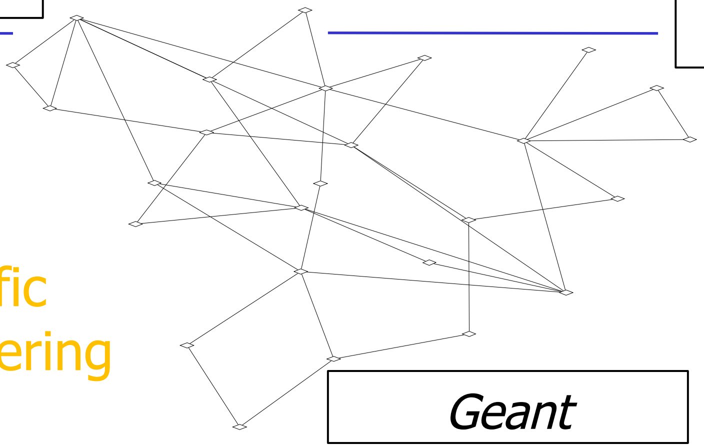
$$\underbrace{0,\ldots,\beta}_{non\ stressed},\ \overbrace{\ldots}^{transient}\ ,\underbrace{\alpha,\ldots,1}_{stressed} \quad \forall p,d \ \ x_j^d(p) \leftarrow x_j^d(p)\times\frac{\alpha}{U(l)}$$

Cartography and traffic matrix
Probing networks - mrinfo and totem tool

Path Diversity

Alternet

Open Transit

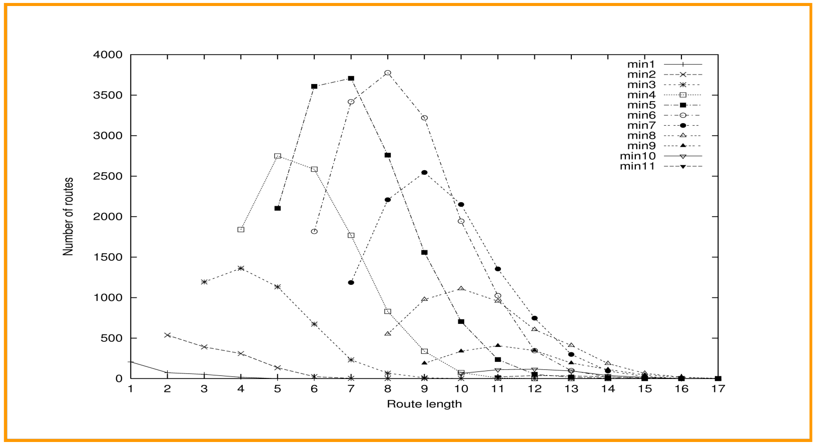


Geant

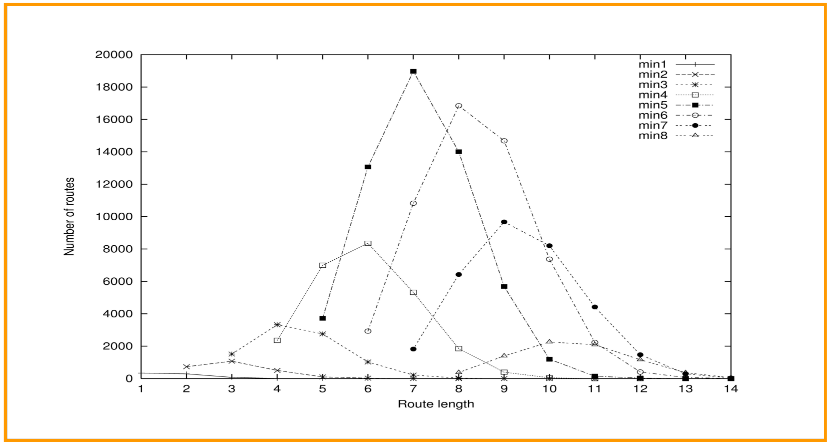
	#node	*#link*	*Diameter*
Alternet	83	334	8
Open Transit	76	206	11
Geant	*23*	*74*	*6*

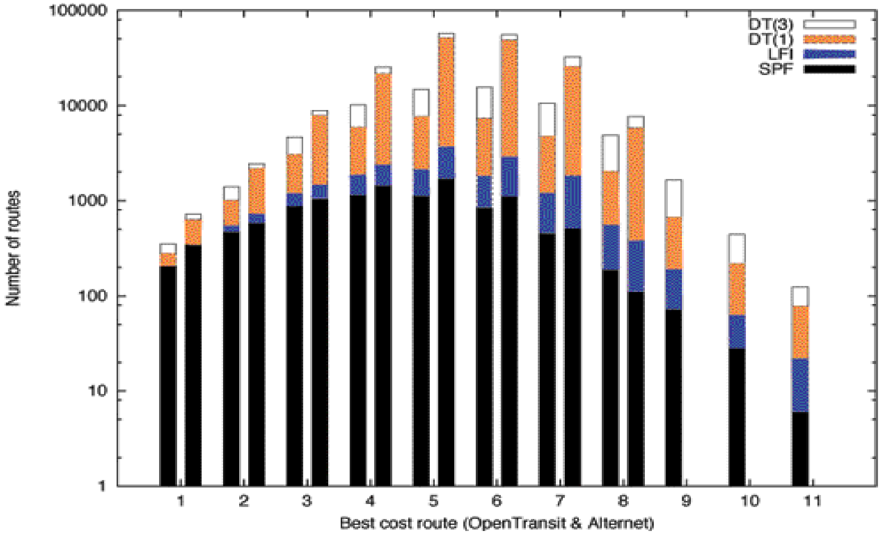
Traffic engineering

Path diversity
Number of routes and rerouting capacities



DT(3)





Open Transit & Alternet (routes number)

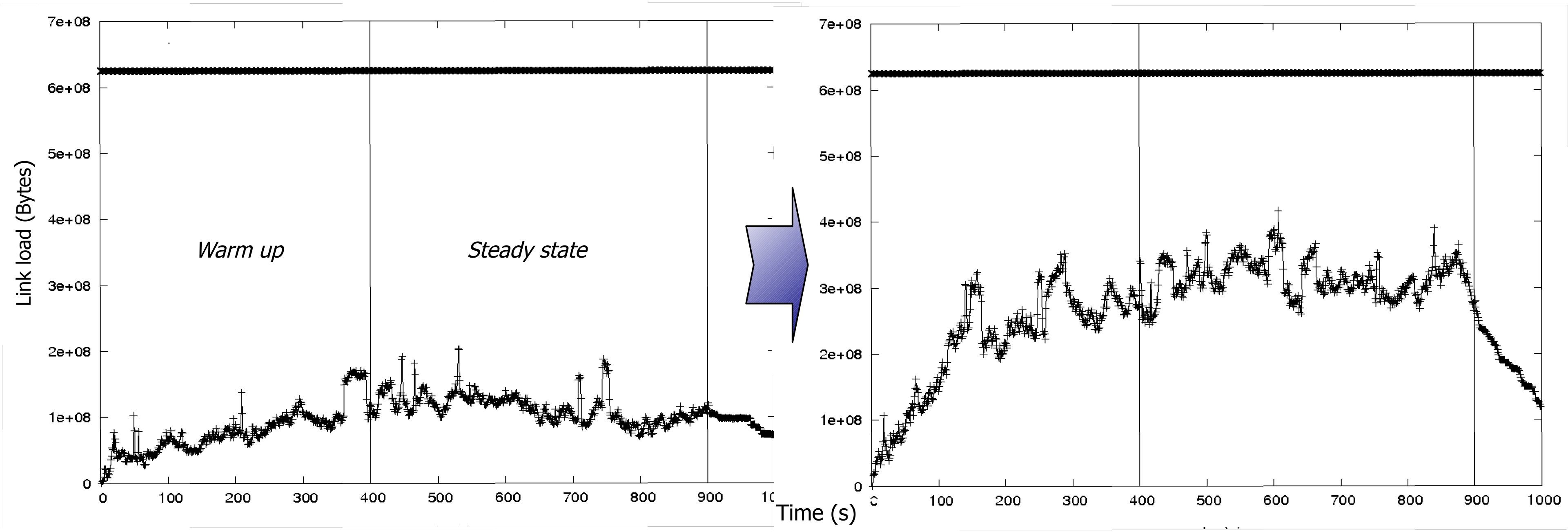
	LFI	*DT(1)*	*DT(3)*
Open transit	18	98	99
Alternet	16	60	78
Geant	37	37	75

Local re-routing capacities

TE results

Simulations setup on Geant network

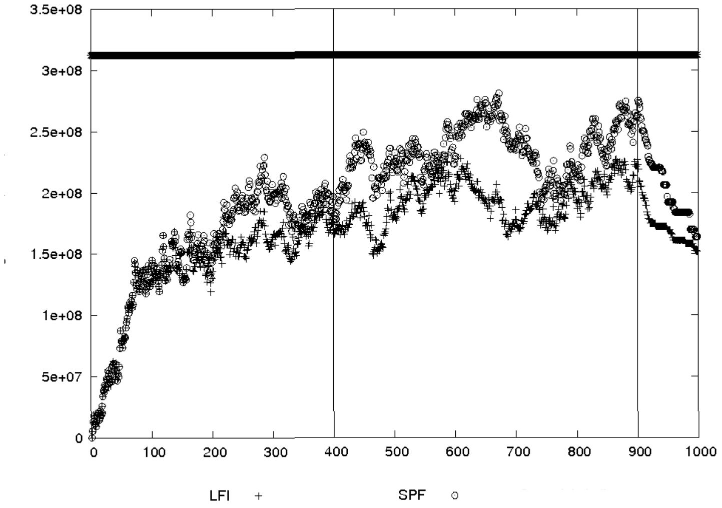
- Based on realistic traces : **Totem** traffic matrix (900 s)
  - *Each entry is decomposed in TCP flows (Reno → Sack)*
  - *GEANT is over-provisioned → artificial congestions*



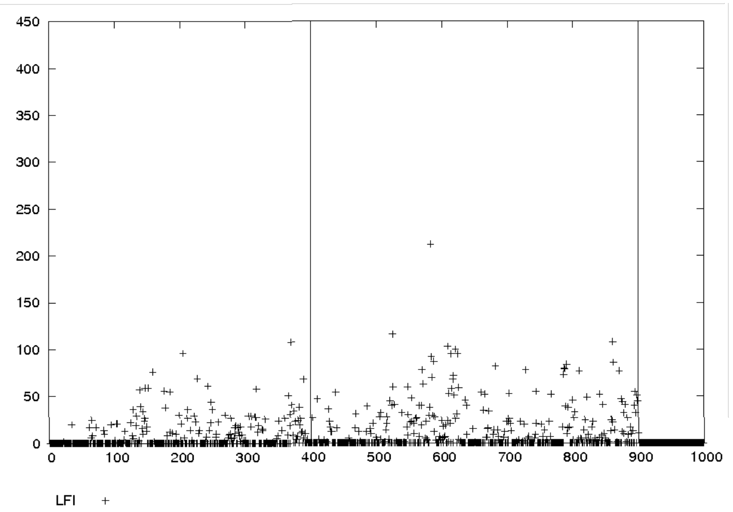
TE results

A single case as an example

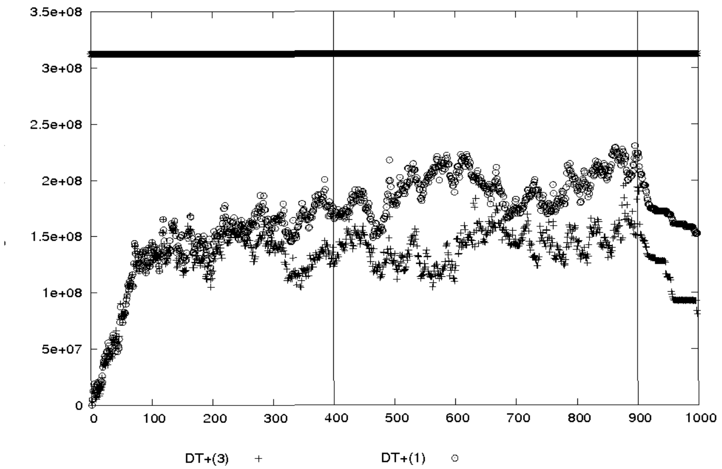
Link load (Bytes)



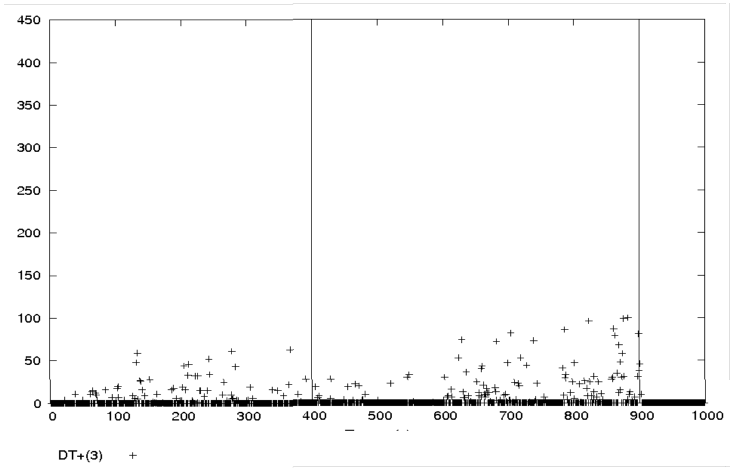
LFI vs SPF



Loss (Dropped packets)

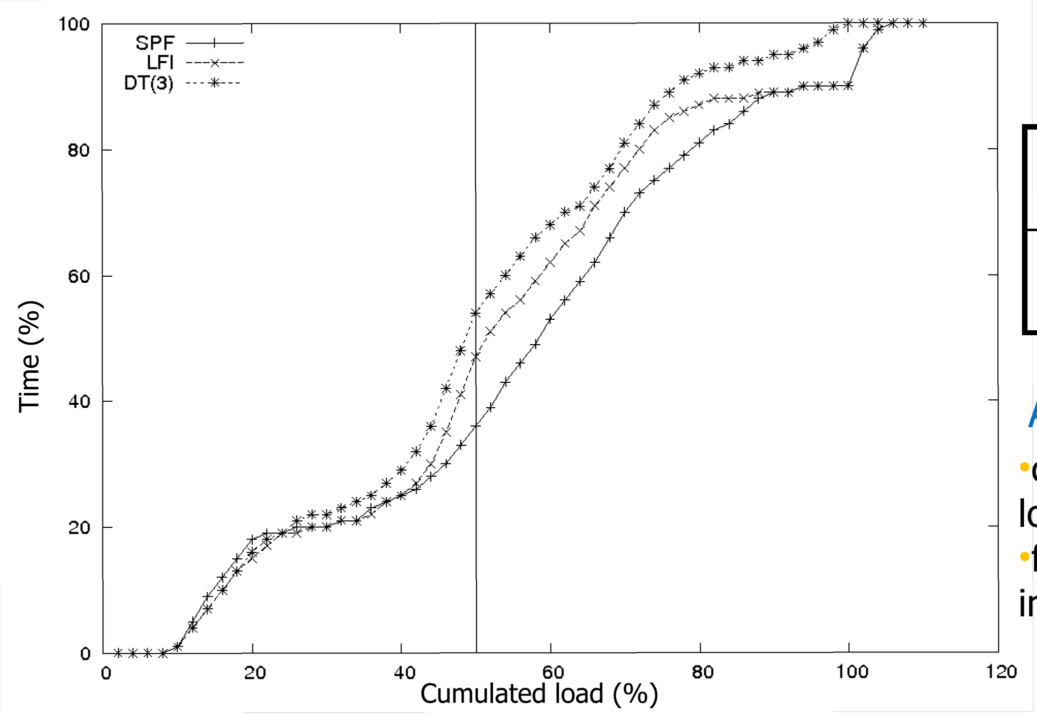


DT(3) vs LFI



TE results
Evaluation

- Configuration parameters and measured indicators
  - *α=0.5, β=0.25 and t=1s (sender windows bounded by 65 packets)*
  - *link load and dropped packets*



Results compilation

	LFI	*DT(1)*	*DT(3)*
Average loss reduction ratio (compared as SPF)	3.8	4.2	6.5
Average load of most loaded link (SPF : 76%)	61.4	61.4	51.8

Averages calculated on 12 simulations :

- congestions are triggered on the most natural loaded link (1→ 1, 1→ n, n→ 1)
- for each run, the link load average confidence intervall (95%) is below 0.1% of the link capacity

Conclusions

Perspectives & work in progress

- Multipath Routing and Path Diversity
  - The efficiency of load balancing scheme depends of the path diversity *(routes number, coverage & cumulative bandwidth)*
  - DT(p)-TE allows to reduce congestions impact
  - Global notification can enhance the redirection coverage
- Current Work
  - Notification and probing protocols
  - Global load balancing problem statement
- Future work
  - Congestions and/or failures scenarios
  - Scalability and inter-domain issues