TEQUILA
Traffic Engineering QUality of service in the Internet at LArge scale

www.ist-tequila.org

Danny Goderis
Alcatel

Tequila Consortium

- **Industrial Partners**
  - **Alcatel, Belgium**
  - **Algosystems S.A., Greece**
  - **France Telecom-R&D, France**
  - **Global Crossing, UK**
- **Universities**
  - **NTUA - National Technical University Athens, Greece**
  - **UCL - University College London, UK**
  - **UniS - The University of Surrey, UK**
- **Research Institutes**
  - **IMEC, Belgium**
  - **TERENA, Netherlands**

TEQUILA Presentations

- **The TEQUILA rationale for QoS delivery**
  - **D. Goderis (Alcatel)**
- **Traffic engineering the multi-service Internet**
  - **G. Pavlou (UniS)**
- **QoS-aware monitoring and measurement**
  - **R. Egan (Global Crossing)**
- **QoS routing over the Internet: a BGP-based approach**
  - **C. Jacquenet (France Telecom)**

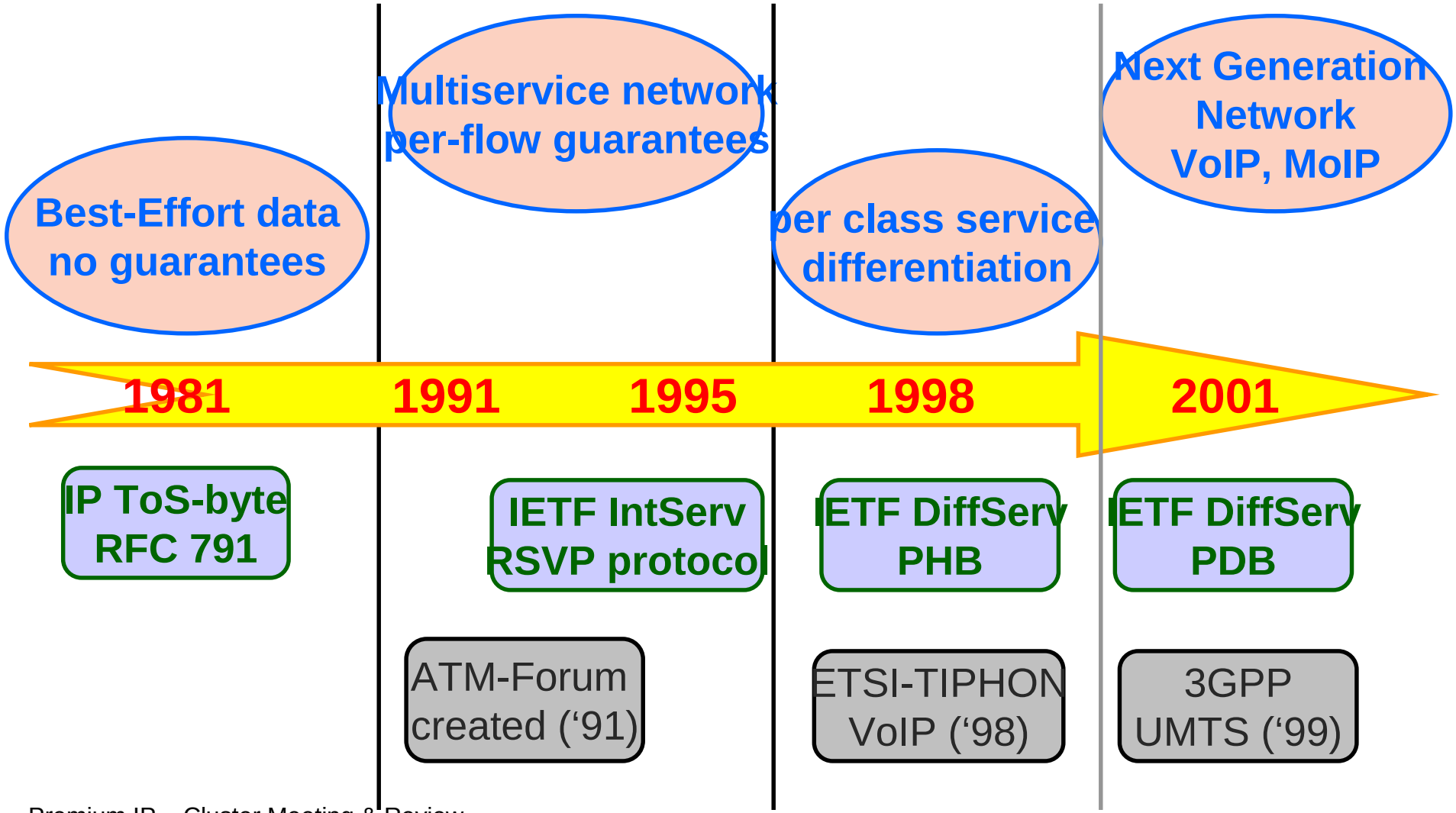
The TEQUILA rationale for QoS delivery

Danny Goderis
Alcatel

Outline

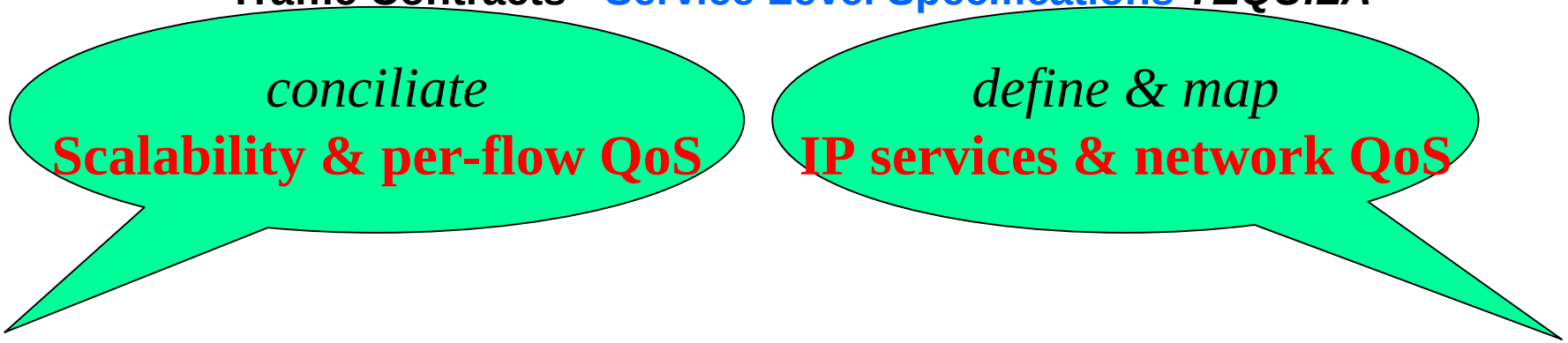
- Introduction
- A DiffServ layered service model
- Service provisioning & admission control
- The TEQUILA model illustrated

The IETF IP QoS Debate...



… and the IETF IP QoS Key Issues

- **IntServ: scalability problem due to per-flow processing & admission**
- **DiffServ tackles scalability by per-class processing**
  - **But only *Edge-to-Edge* guarantees for *aggregate* packet streams...**
    - **no hard *per-flow* guarantees**
  - **...and missing standards**
    - **Traffic Contracts - Service Level Specifications *TEQUILA***

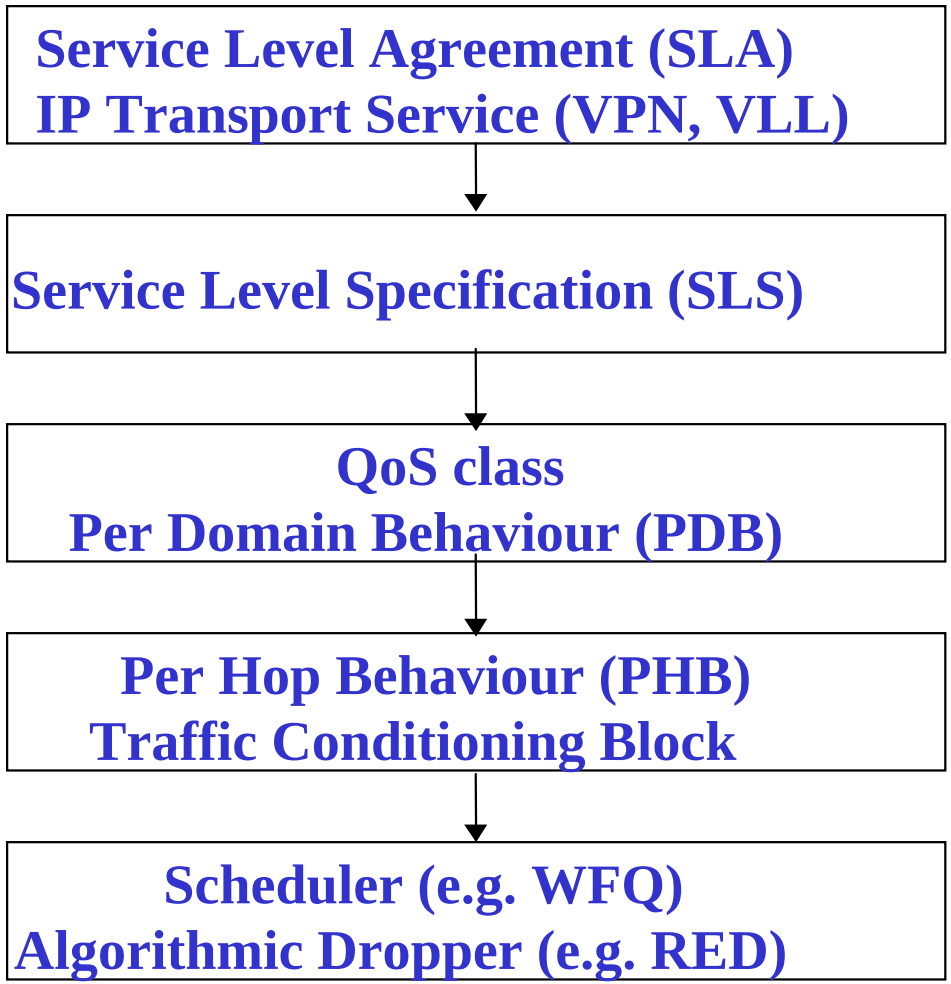


TEQUILA Key Concepts

	PSTN	*TEQUILA*
Technology	Circuit-switching	IP DiffServ
Granularity	64 kbps	DSCP, PHB
Service	Voice call	Service Level Specifications
Dimensioning	Erlang-B	Resource Provisioning Cycle
Allocation	CAC	2-level Admission Control

Part 2
A DiffServ Layered Service Model

From SLA to Packets



- Non-technical terms & conditions
- technical parameters :{SLS}-set

- IP service traffic characteristics
- offered network QoS guarantees

- Network QoS capabilities
- DiffServ edge-to-edge aggregates

- Generic Router QoS capabilities
- DiffServ core & edge routers

- vendor & product specific implementation

TEQUILA SLSs

Parameter Group	*Description*
Customer-user Id	Identifies the *customer*
Flow descriptor	*Packet stream* (DSCP, IP addresses, etc)
Service Scope	*Geographical region* (ingress–egress)
Service Schedule	Specifies *when* the contract is applicable
Traffic descriptor	*Traffic envelop* (e.g. a token bucket)
QoS Parameters	*QoS guarantees* (delay, jitter, packet loss)
Excess Treatment	*Traffic conditioning* (dropping, remarking)

Tequila QoS Classes ~ PDBs

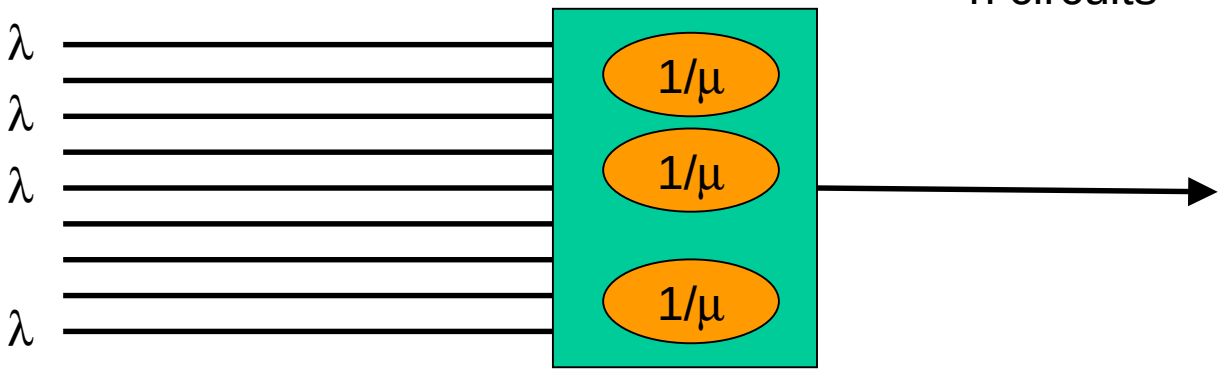
- **QoS class = [OA | delay | loss]**
  - **Ordered Aggregate ~ PHB scheduling class**
    - **EF, AFx, BE**
  - **delay**
    - **edge-to-edge maximum delay**
    - **worst case or probabilistic (percentile)**
    - **delay classes (min-max intervals)**
  - **loss**
    - **edge-to-edge packet loss**
    - **probability**

Part 3
Service Provisioning & Admission Control

PSTN Dimensioning

N subscribers
BHCA = λ, MCD = 1/μ

Central Office
n circuits

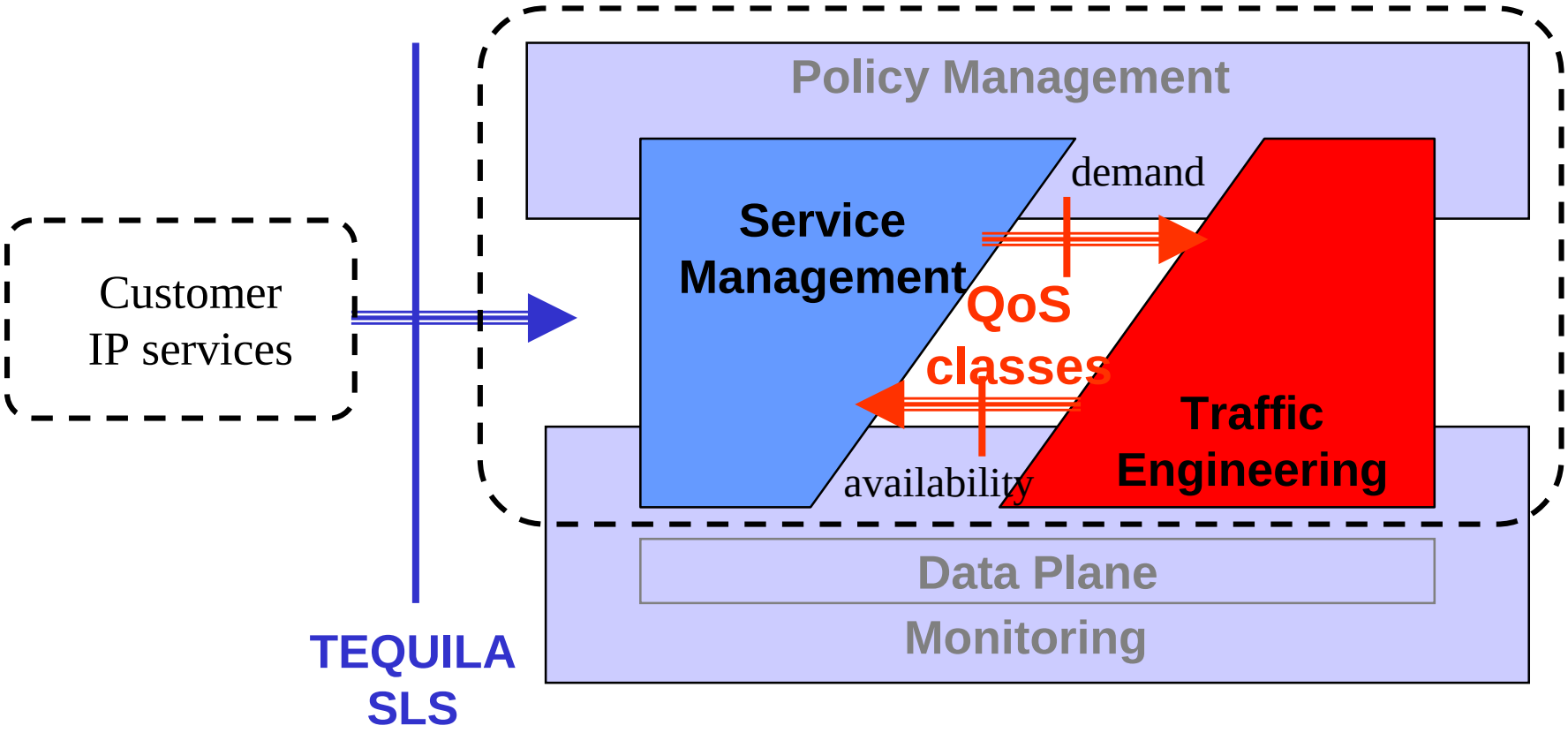


Erlang B

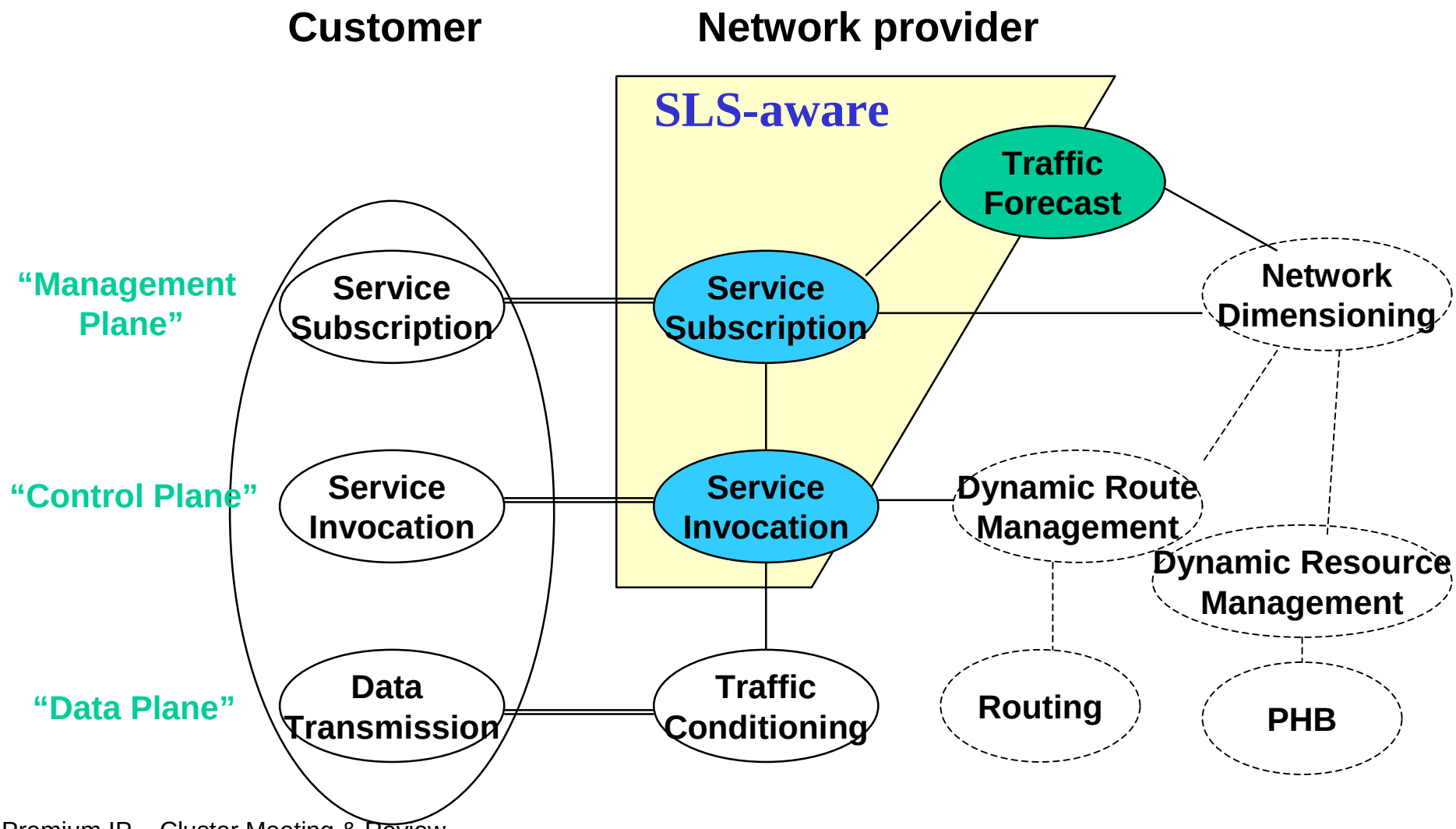
Call blocking probability = $\dfrac{\dfrac{\rho^n}{n!}}{\sum_{i=0}^{n}\dfrac{\rho^i}{i!}} \leq \varepsilon$

ρ = Nλ/μ

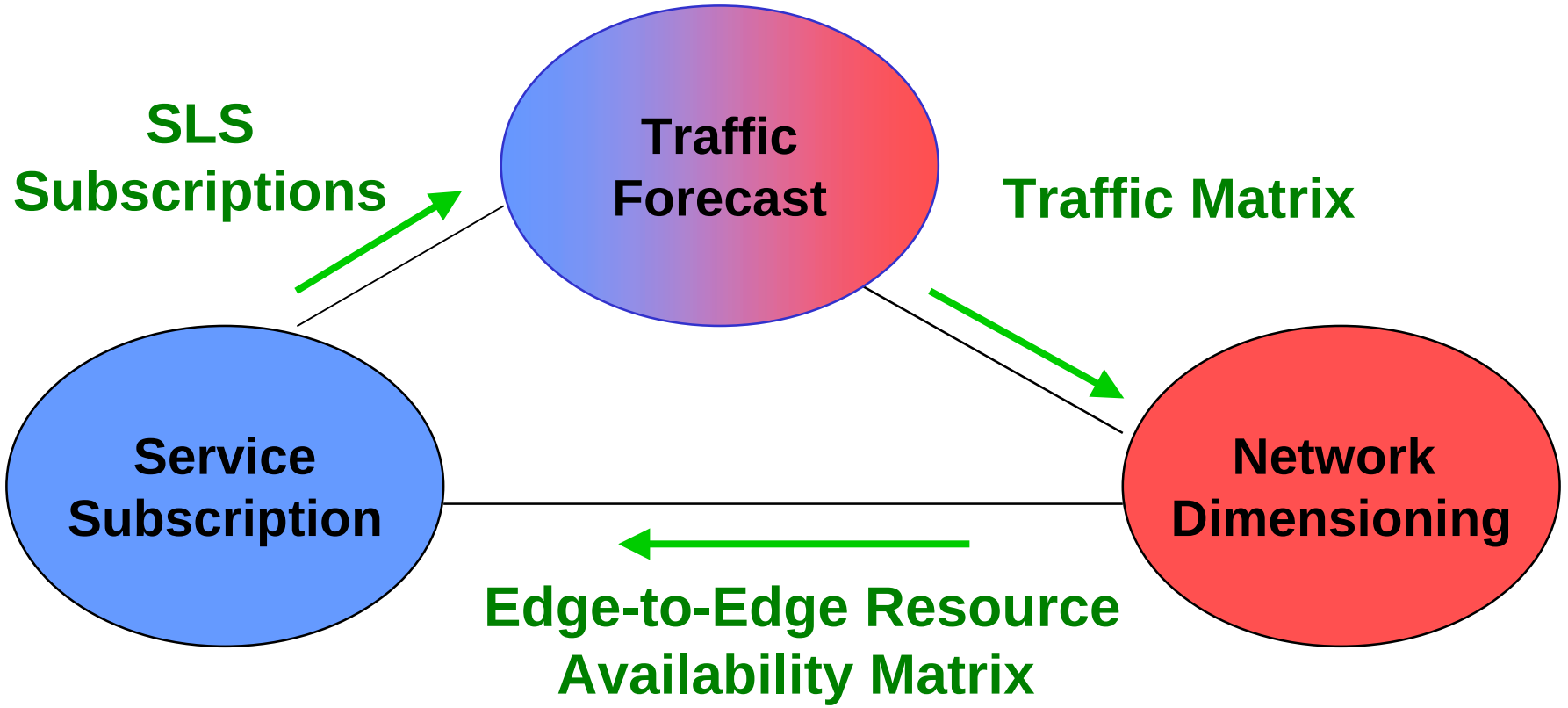
Tequila Approach for IP QoS Delivery



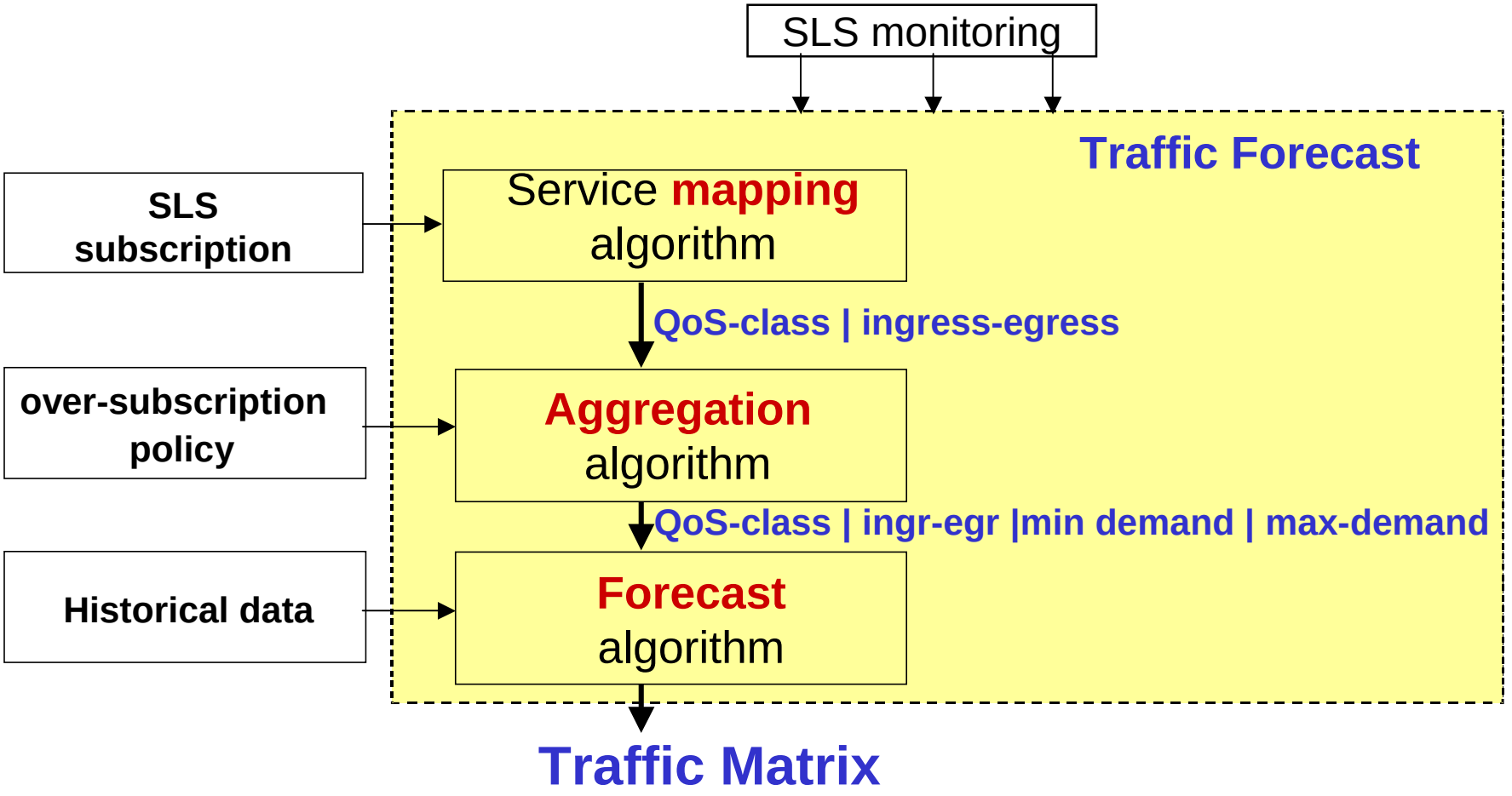
Service Management



Resource Provisioning Cycle

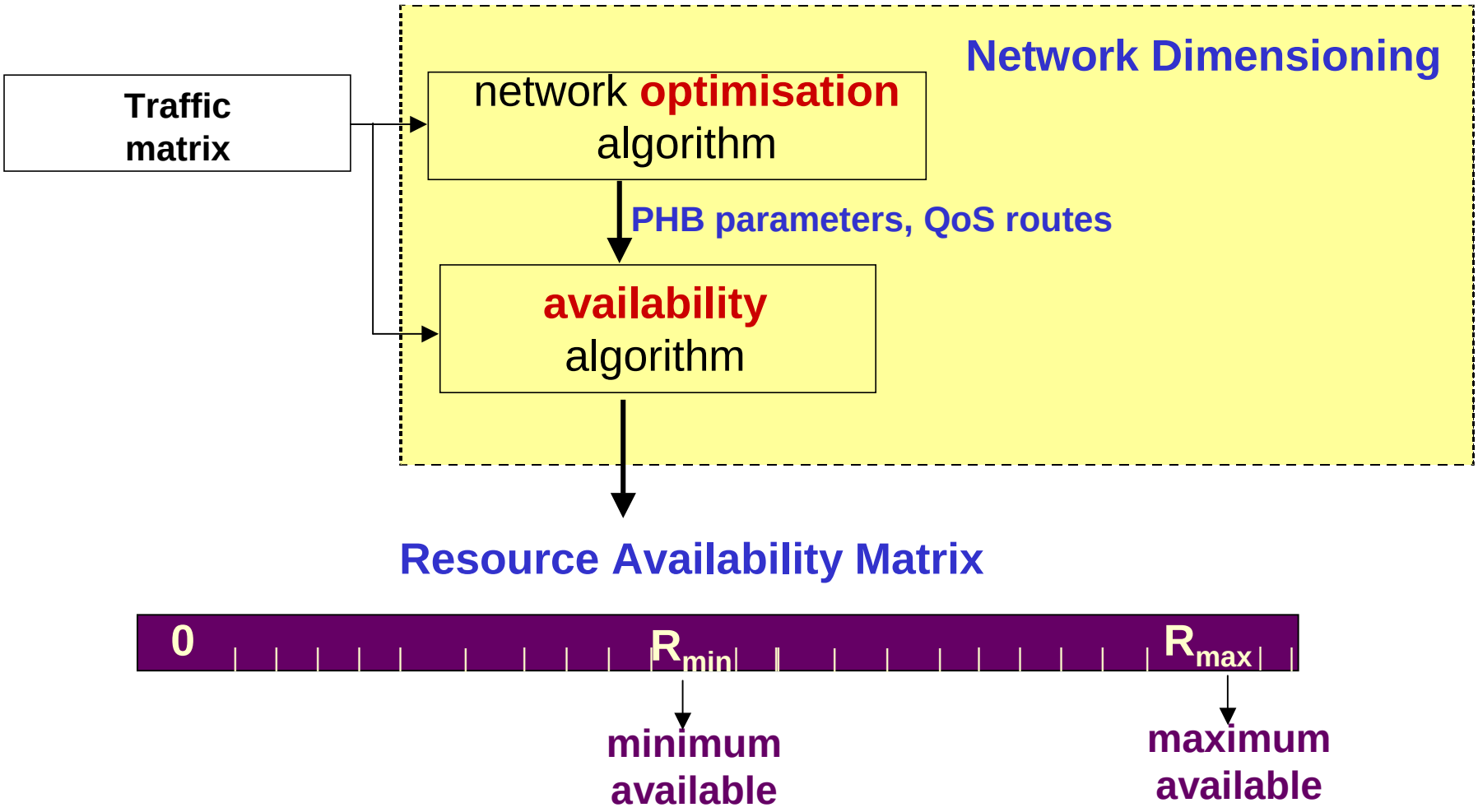


Estimating the Traffic Matrix

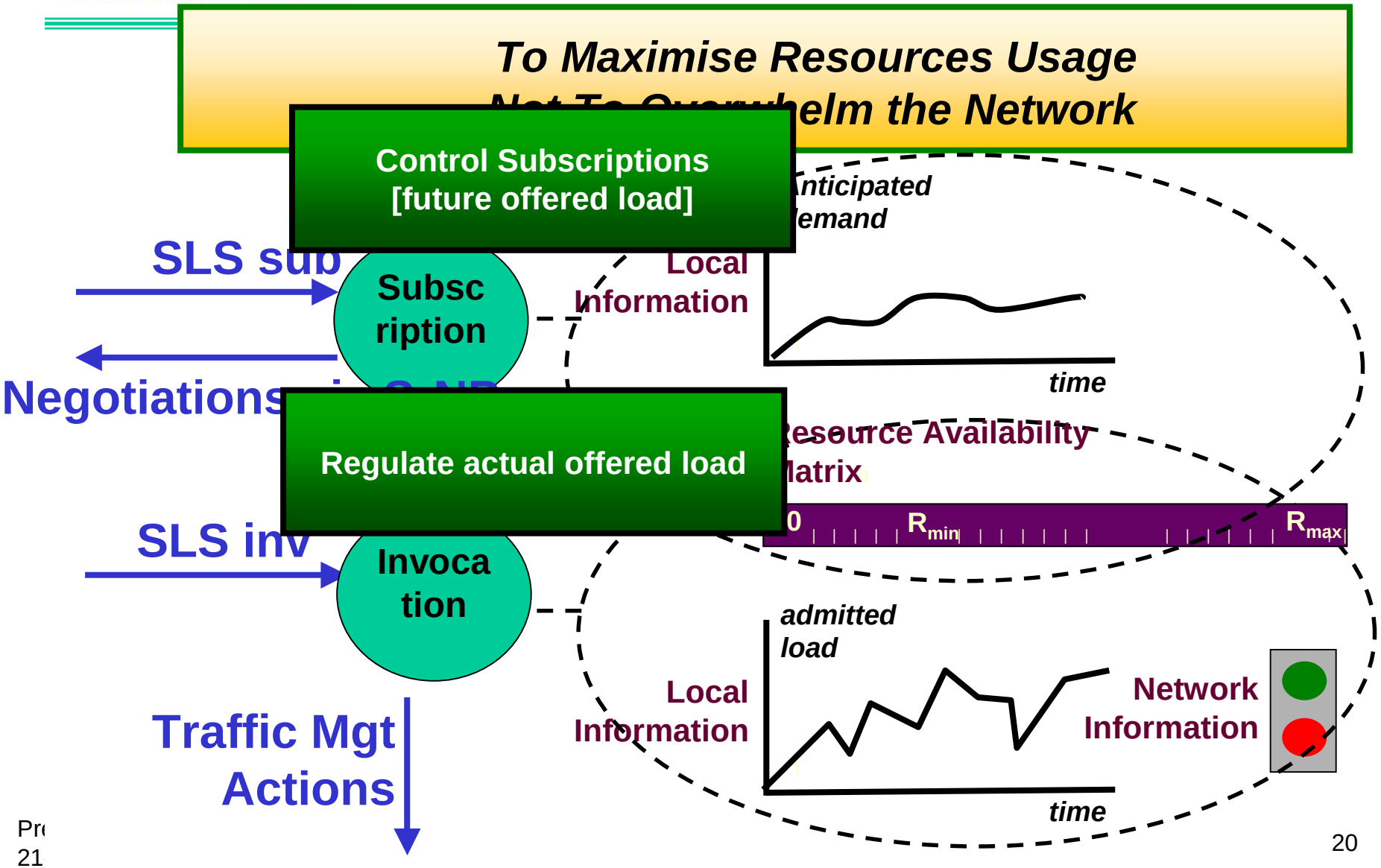


[QoS class | ingress-egress | min-demand - max-demand]

Generating the Resource Availability Matrix

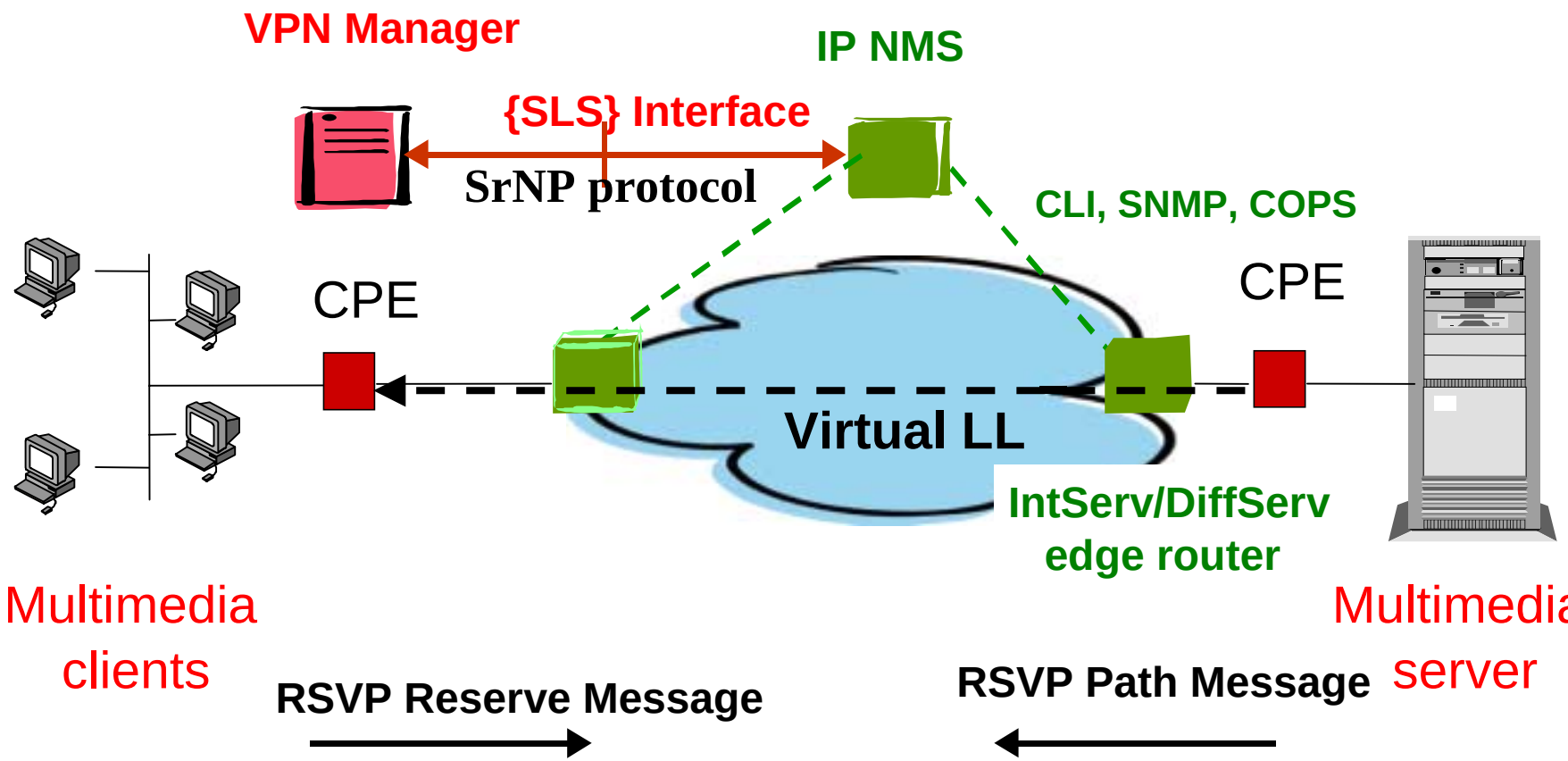


Two-level Admission Control

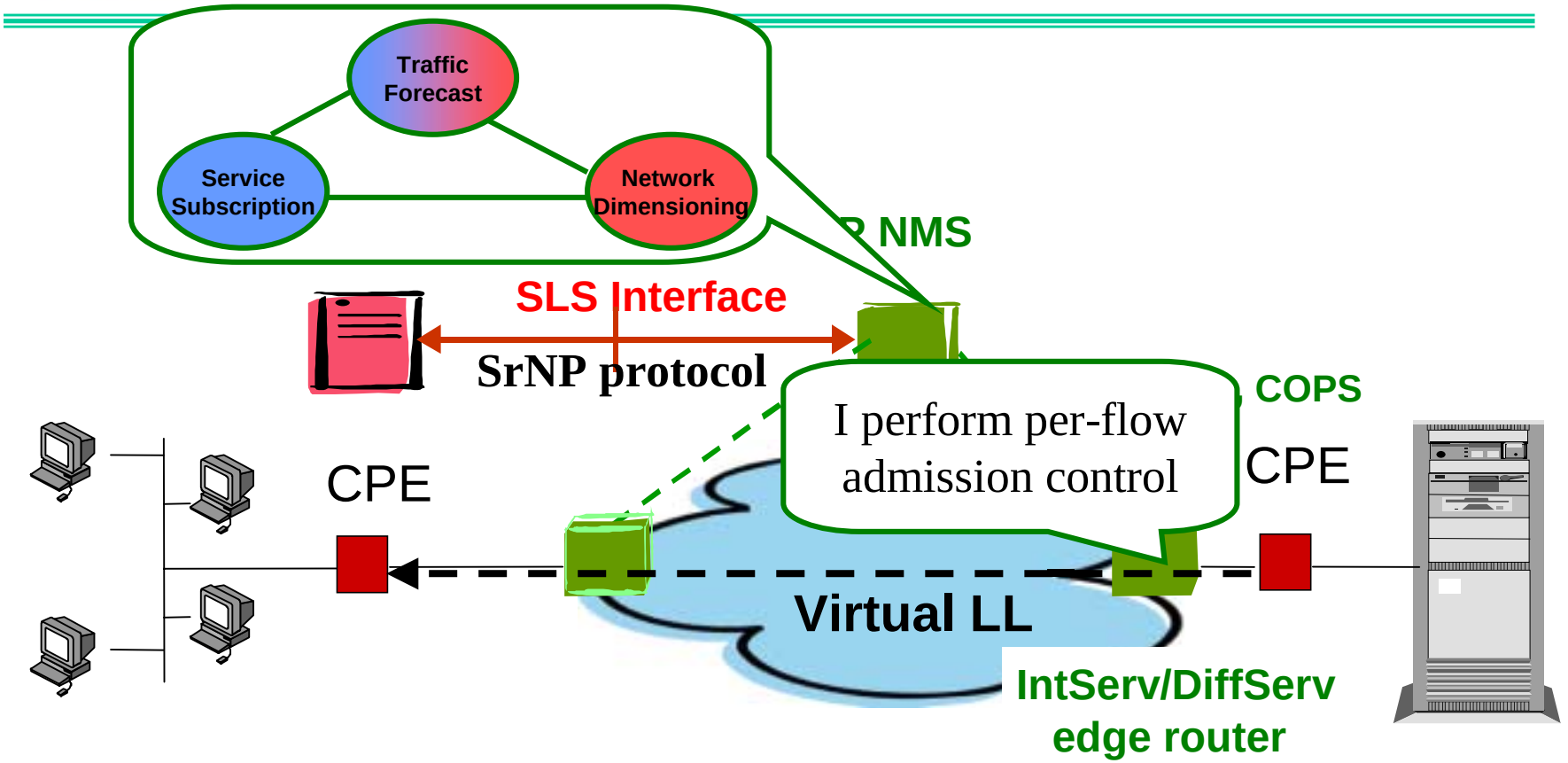


Part 4
The TEQUILA Model Illustrated

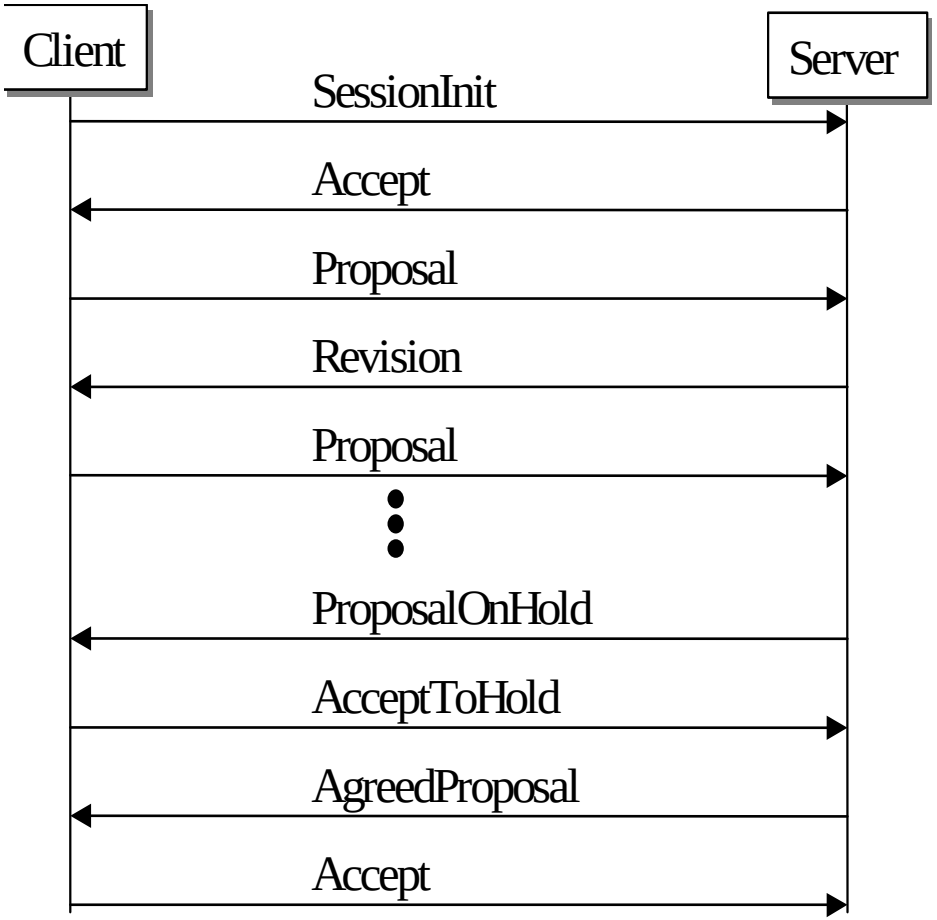
Multiplexing Multimedia in a VLL



Multiplexing Multimedia in a VLL



Service Negotiation Protocol - SrNP

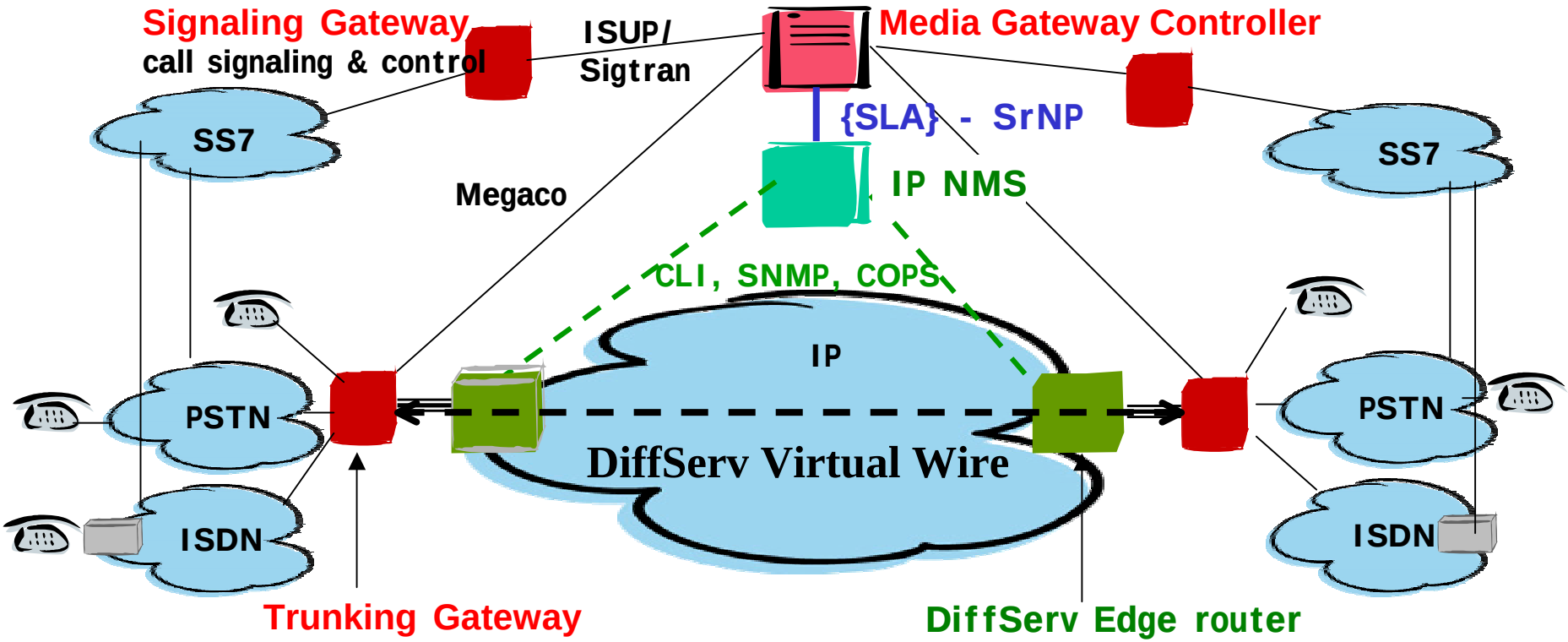


- **Client-server based**
- **Form-fill oriented**
- **Messaging is content-independent**
- **Protocol stacks**

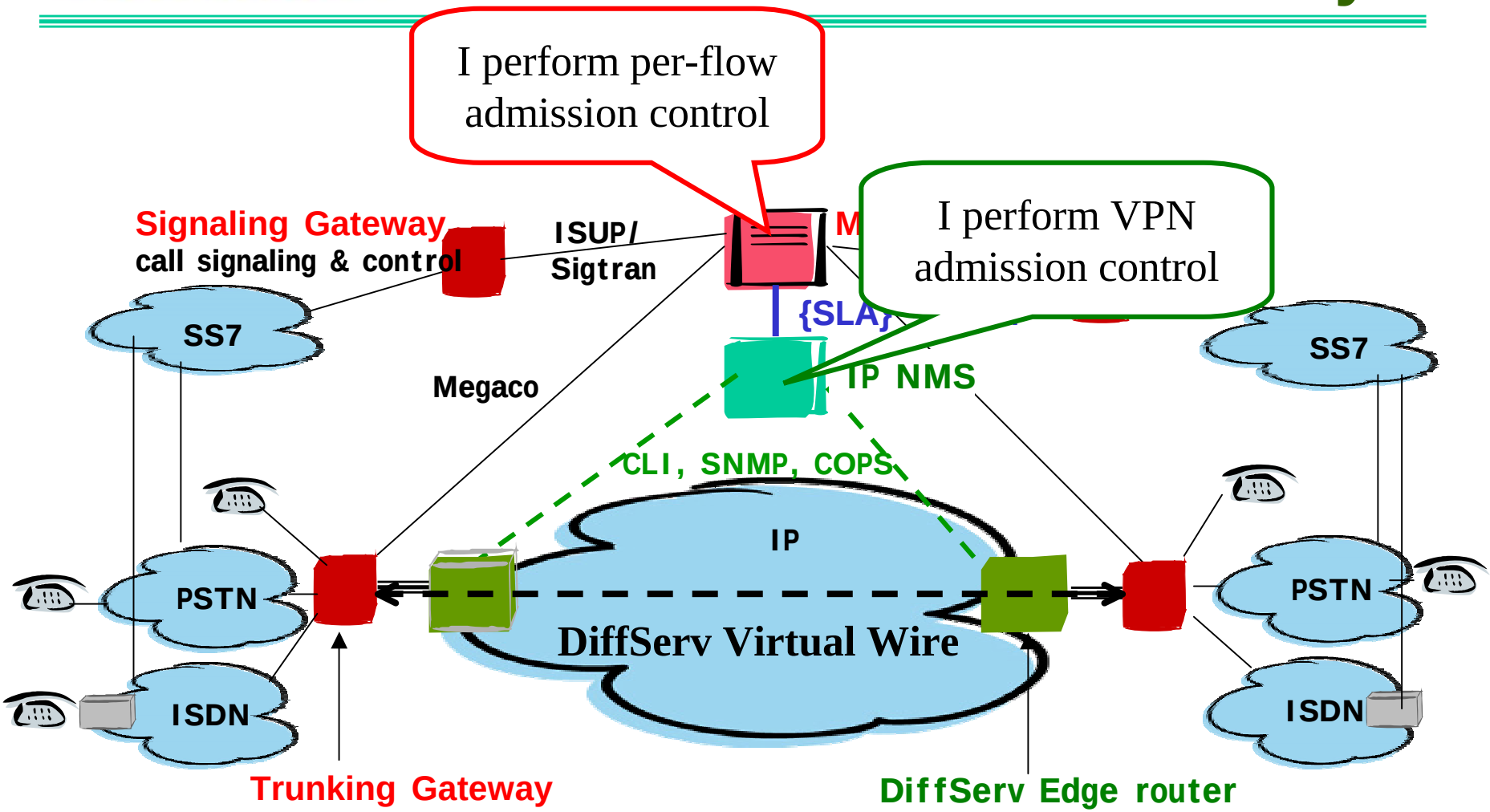
SrNP
XML
HTTP,SMTP,IIOP
TCP/IP

SrNP
TCP/IP

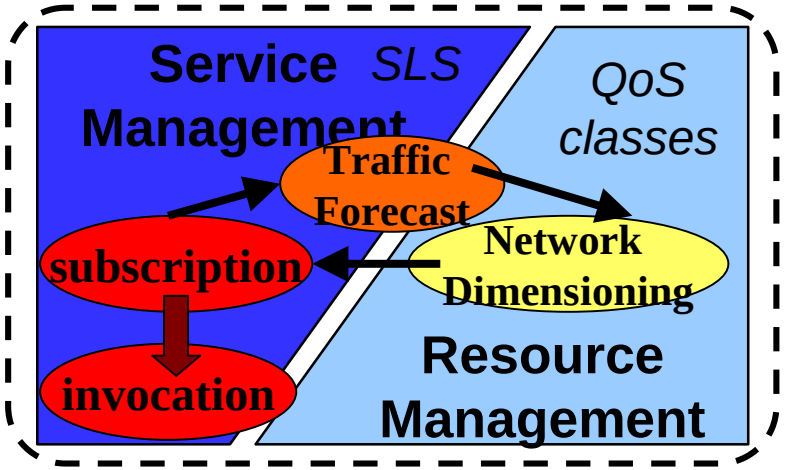
Connecting Trunking Gateways



Connecting Trunking Gateways



Conclusions



- **Clear separation of service & resource management**
  - **service system: only *edge-to-edge* view on the network**
  - **resource system: only QoS class aware (*no SLS-awareness*)**
- **Two-level admission control**
  - **long-term IP aggregates based on *resource provisioning cycle***
  - **short-term flows based on long-term guidelines**

Traffic Engineering the Multi-Service Internet

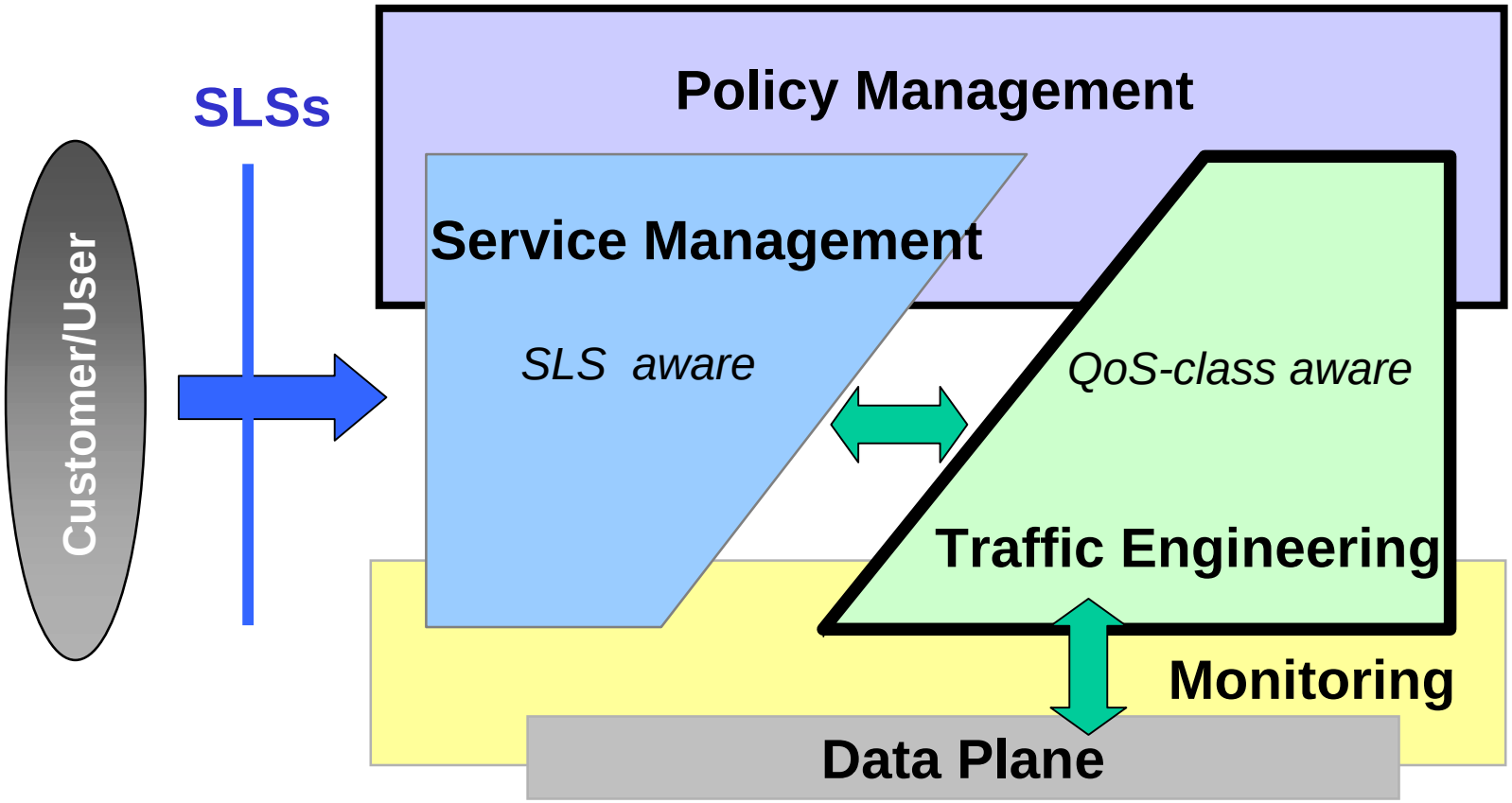
Prof. George Pavlou
Centre for Communication Systems Research
University of Surrey, UK
G.Pavlou@eim.surrey.ac.uk

TEQUILA Consortium

Introduction

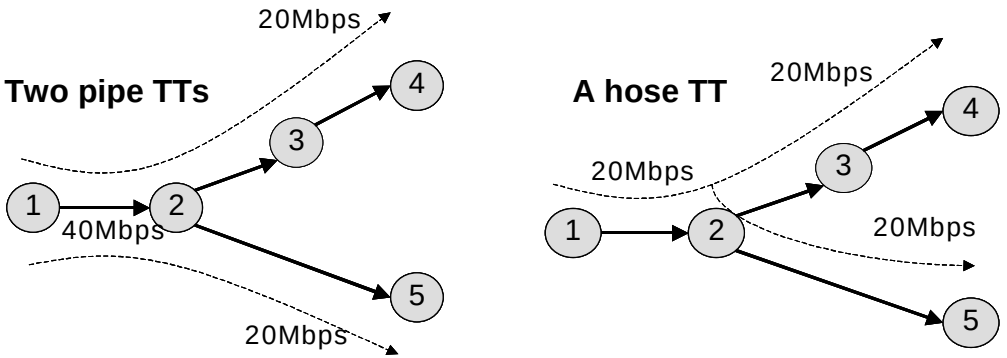
- Internet: the global multi-service network
- Need for scalable Quality of Service (QoS) solutions
- Differentiated Services (DiffServ)
  - Classify, mark and police at the edges
  - Specified “per-hop behaviours” (PHBs) to traffic aggregates
  - Scalability compared to per-flow reservation approaches
- Traffic Engineering
  - Control the manner traffic is mapped to and treated by network to achieve specific performance objectives
  - Of paramount importance in DiffServ since there are no explicit resource reservations to flows within the network
- Key problem: how to traffic engineer a DiffServ domain to meet edge-to-edge QoS requirements as dictated by contracted SLSs

Functional Model for QoS

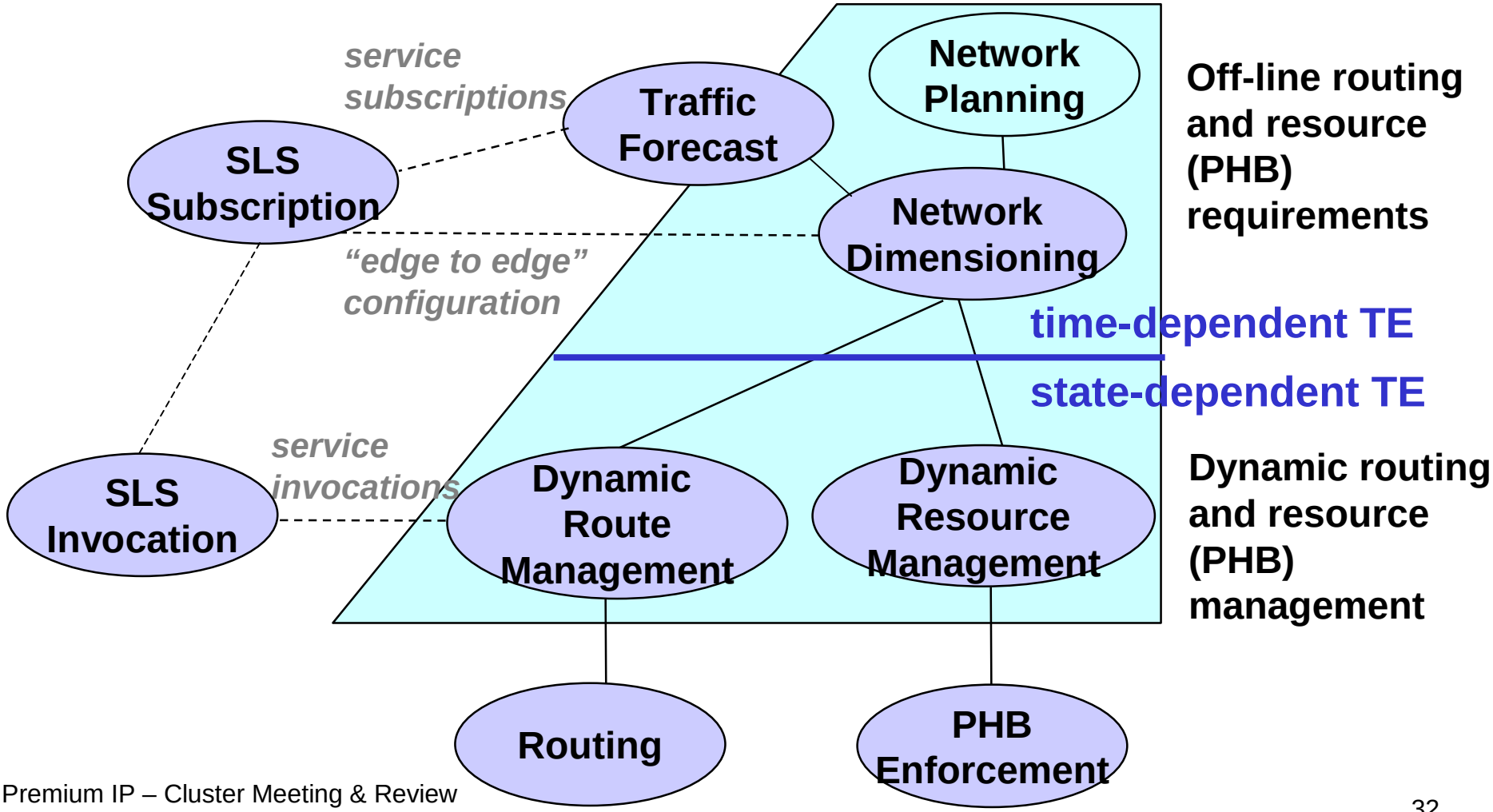


Traffic Model

- **Traffic Trunk (TT)**
  - **Ingress, set of egresses**
    - **pipe (1:1) or hose (1:N) model**
  - **General case the hose model**
    - **results in a tree with logically associated “tree bandwidth”**



Traffic Engineering Model



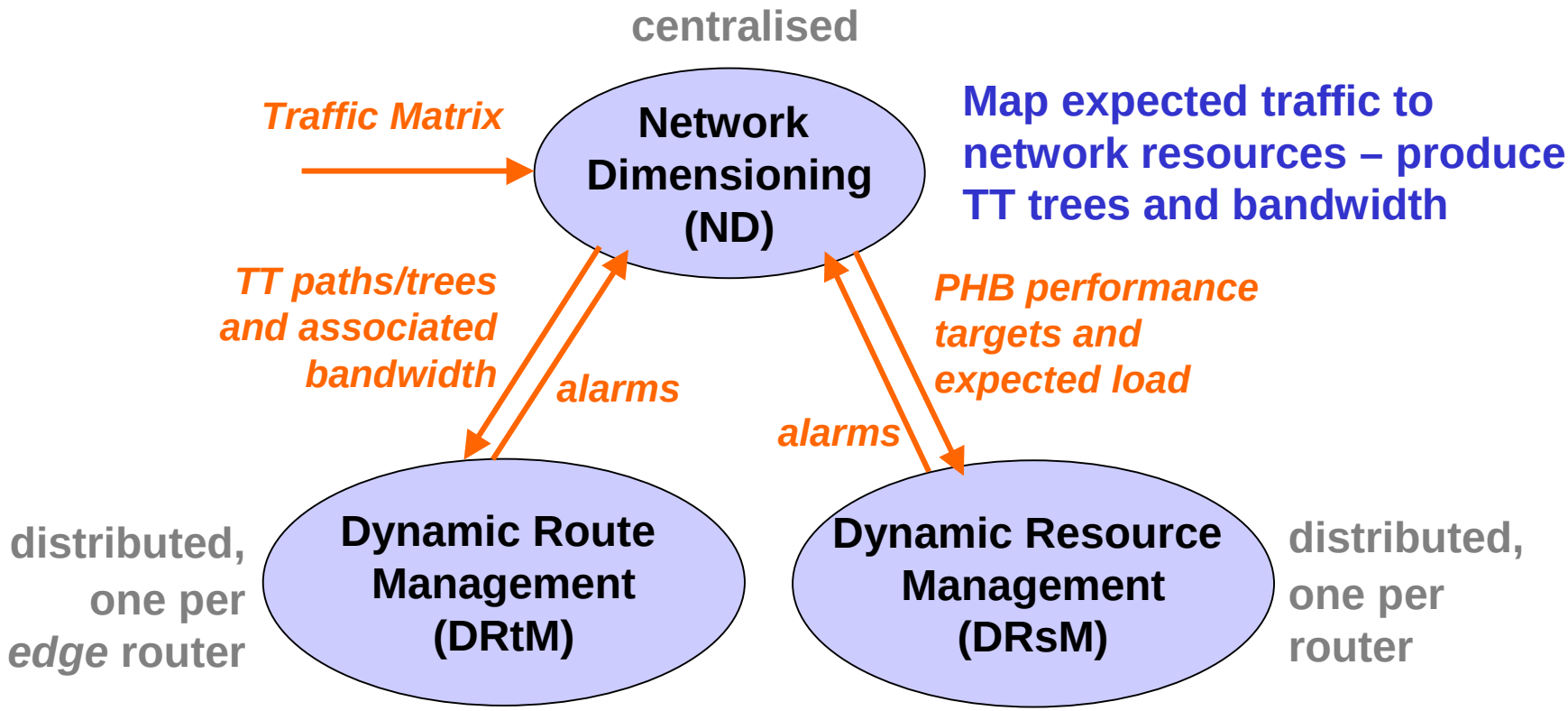
Traffic Forecast

- **Traffic Forecast is the "glue" between the customer-oriented (SM) and resource-oriented (TE) parts**
- **Estimates expected traffic demand (Traffic Matrix)**
  - **Derived from service contracts (SLSs) and other information (SLS usage, business policies, forecast/projection)**
  - **A traffic matrix for every provisioning cycle**
- ***Aggregation* required for scalability**
  - **Maximum entries per edge node $N*2^{N-1}*Q$ for N edge nodes and Q QoS classes**

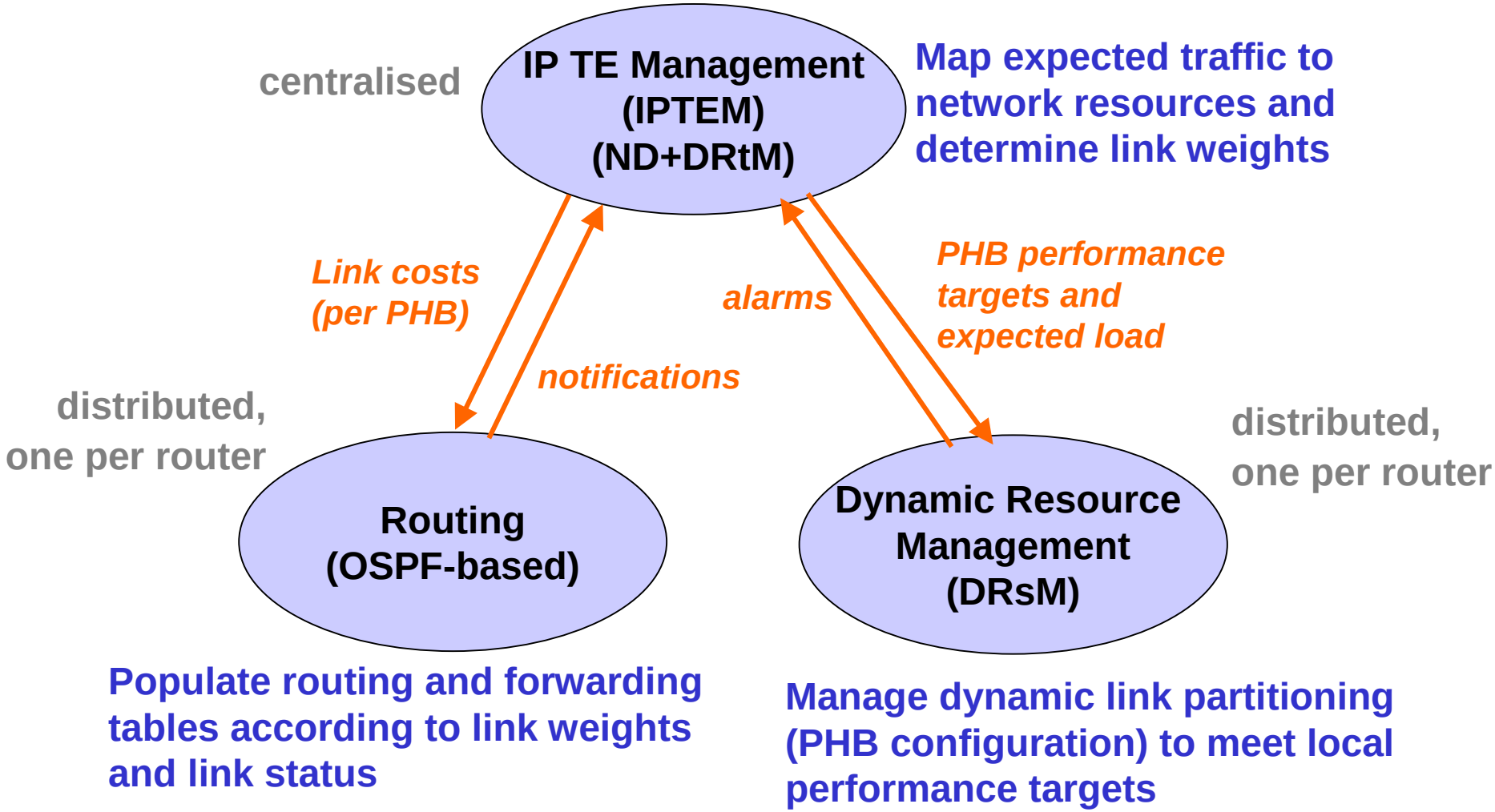
TE Approaches

- **MPLS-based**
  - **Explicit routing through LSPs with alternative LSPs for source-destination combinations for load balancing**
  - **LSPs are *not* explicitly associated with bandwidth**
- **IP-based**
  - **Hop-by-hop routing, OSPF-based with Equal Cost Multi Path (ECMP) for load balancing**
  - **Assignment of link-weights**
- **Loss and delay constraints are translated to route hop-count constraints (PHBs are associated with delay and loss bounds)**

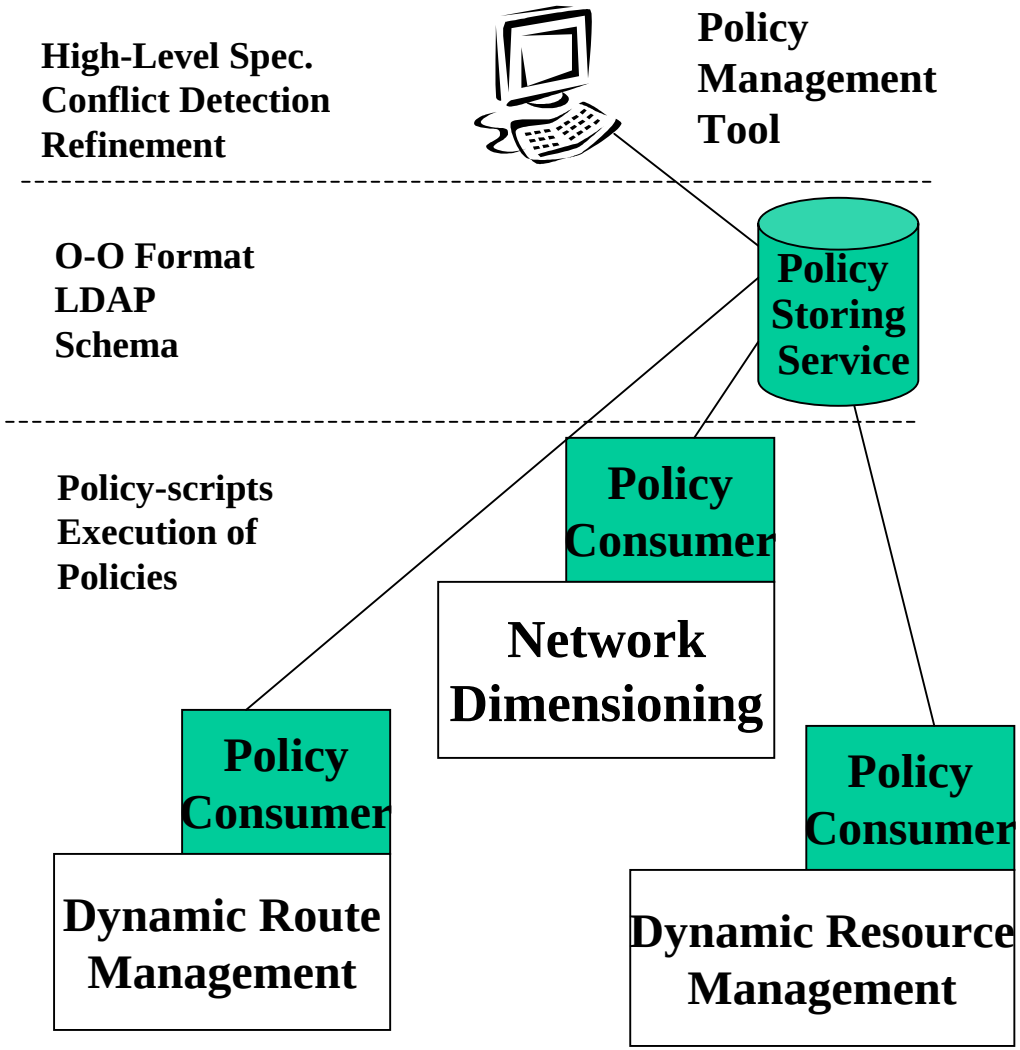
MPLS-based TE



IP-based TE


centralised
IP TE Management (IPTEM) (ND+DRtM)
Map expected traffic to network resources and determine link weights
Link costs (per PHB)
notifications
alarms
PHB performance targets and expected load
distributed, one per router
Routing (OSPF-based)
Dynamic Resource Management (DRsM)
distributed, one per router
Populate routing and forwarding tables according to link weights and link status
Manage dynamic link partitioning (PHB configuration) to meet local performance targets

Policy-based TE - Hierarchical Approach



High-level Policies may result in the introduction of related policies at lower layers, mirroring the system's hierarchy:

ND Policies

DRtM/DRsM Policies

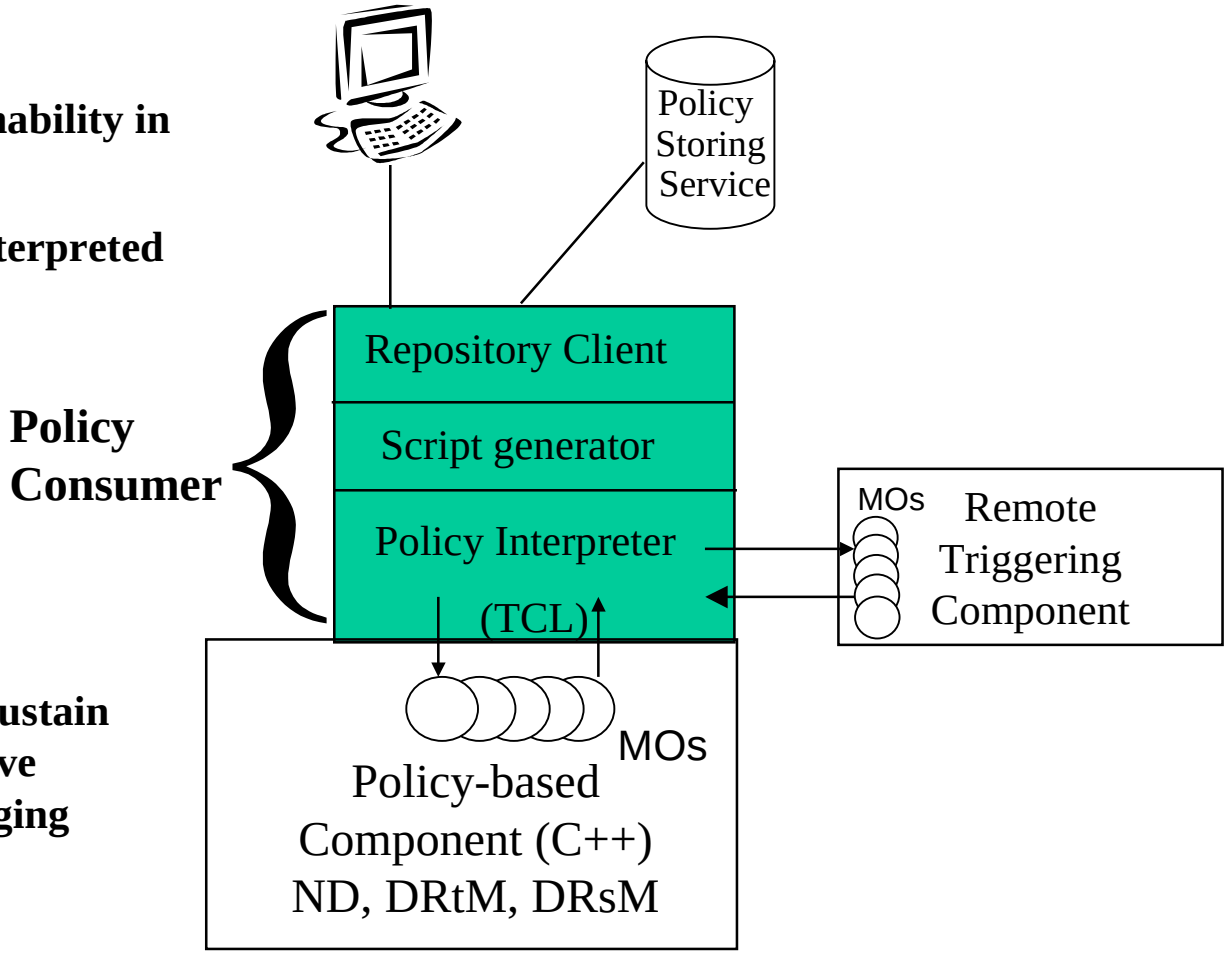
Policy Consumer Decomposition

Policies are seen as:

- **a means to achieve programmability in the Tequila System**
- **logic (scripts) downloaded, interpreted and executed on the fly**

Target:

- **a TEQUILA system able to sustain requirement changes and evolve through policies without changing its initial "hard-wired" logic**



Network Dimensioning Policies

- **Setting initial ND parameters e.g. maximum number of alternative trees, cost function, dimensioning period**

- **Influencing the capacity allocation and LSP creation e.g. allocation of network/link allocated bandwidth per class of service, explicit setup of LSPs, treatment of spare/over-provisioned capacity**

Network Dimensioning Optimisation Problem (MPLS)

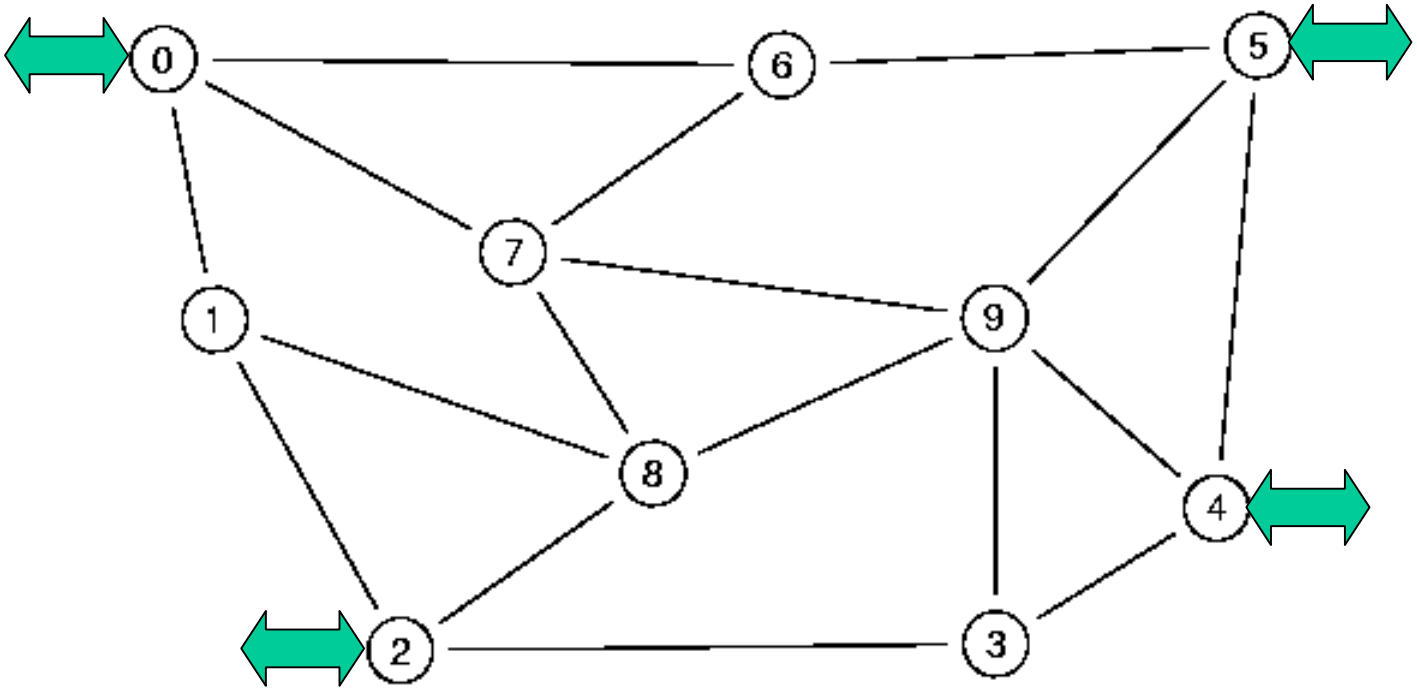
- **Satisfy the QoS requirements of traffic trunks**
  - **Avoid overloading parts of the network (I)**
  - **Minimise overall network utilisation (II)**
- **Assuming a cost function *f(x)* per PHB depending on the current load and *F(x)* the derived cost function per link**
  - **minimise** $(\max_{l \in E} F_l)$ **satisfies (I)**
  - **minimise** $\sum_{l \in E} F_l$ **satisfies (II)**
- **Combined objective function**
  - **minimise** $\sum_{l \in E} (F_l)^n$ **satisfies both: (II) for n=1 and (I) for n-> ∞**

Optimisation Approach

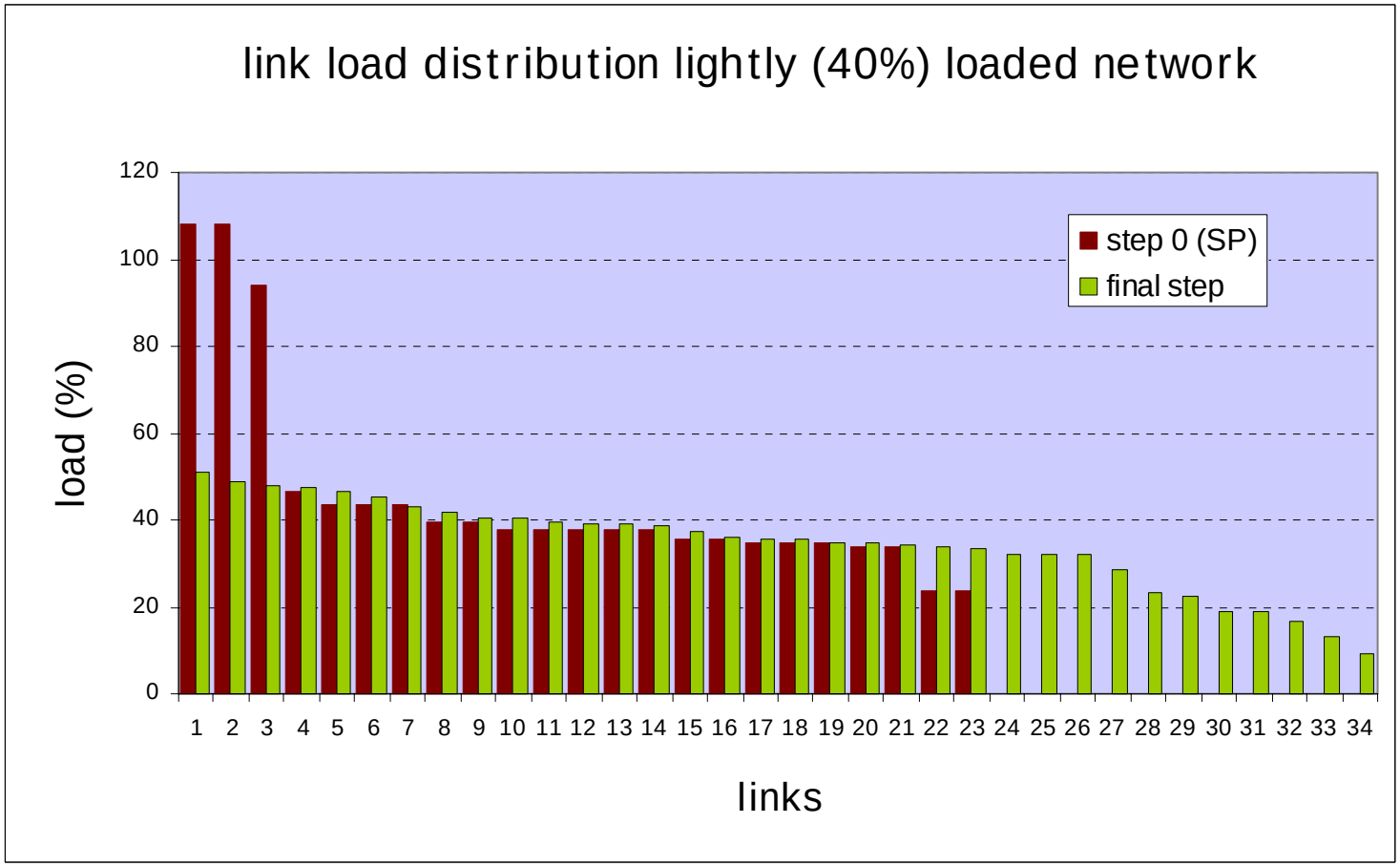
- **Solution-based on an iterative procedure**
  - **Start with an initial TT allocation**
  - **Improve gradually on this solution by moving capacity to “better” paths**
- **At each step, a constrained shortest path algorithm is used for pipe trunks and a constrained Steiner tree algorithm for hose trunks**
  - **The algorithms are constrained due to delay & loss bounds, translated to a hop count constraint**
- **Iterative algorithm works well and converges quickly for pipe trunks**
  - **Results are presented next**
- **The constrained Steiner tree algorithm also works well, in the next step we will be integrating the two**

Example Network

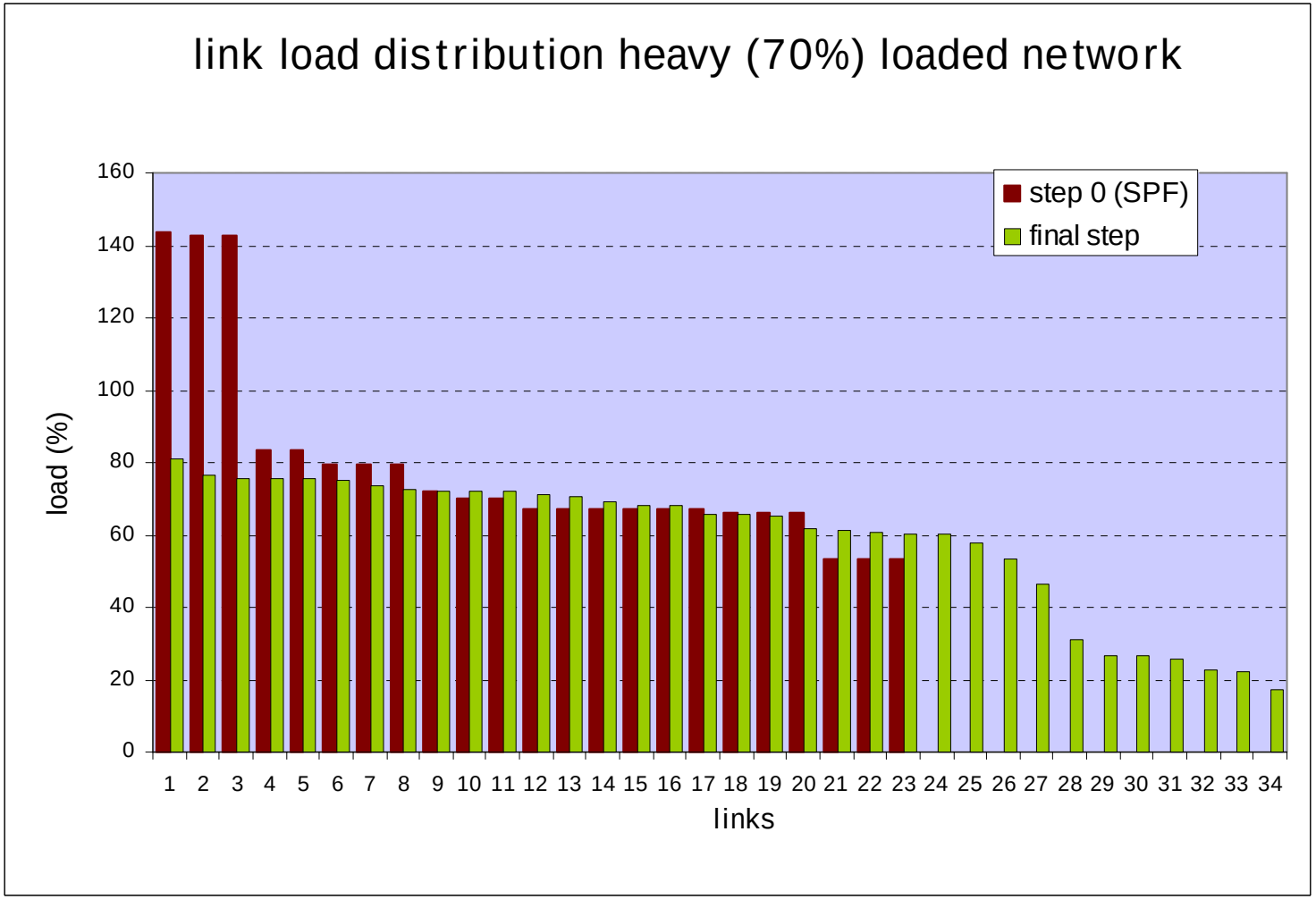
- **10 node network – 4 edge nodes { 0, 2, 4, 5 }**



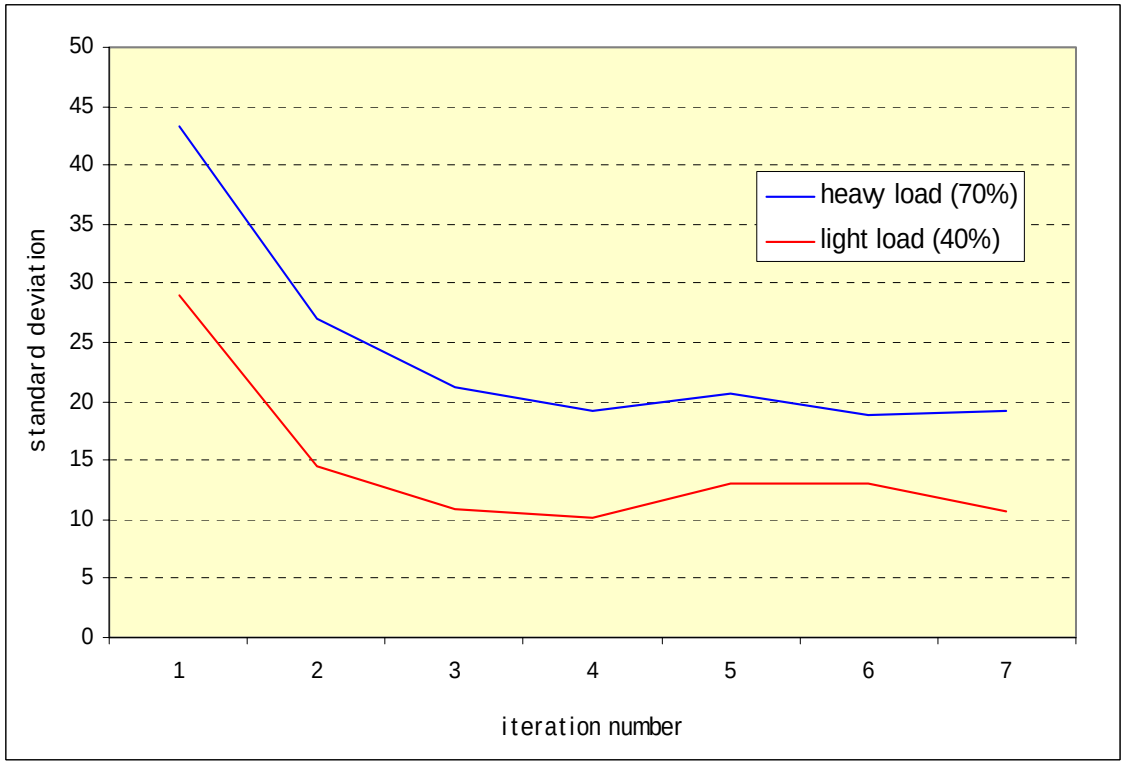
Link Load Distribution – light load (40%)



Link Load Distribution – heavy load (70%)

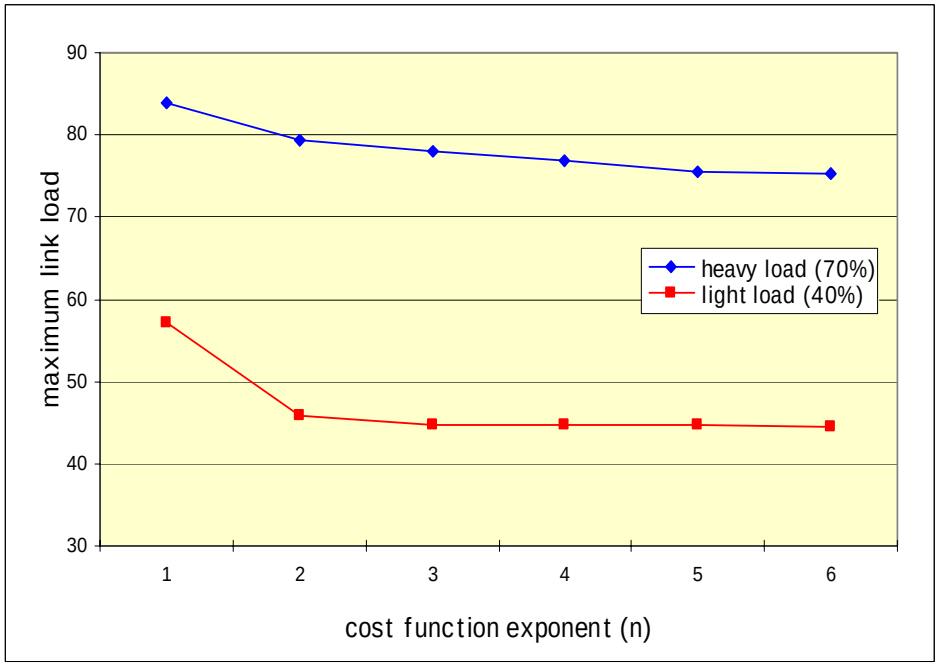


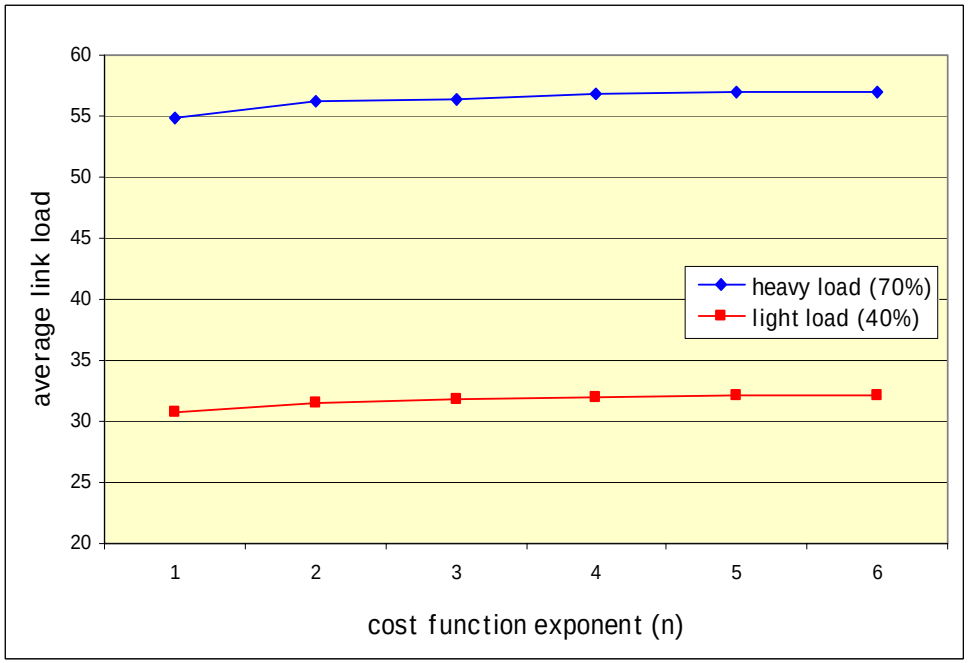
Stepwise (Per-Iteration) Solution Improvement



cost function exponent n=2

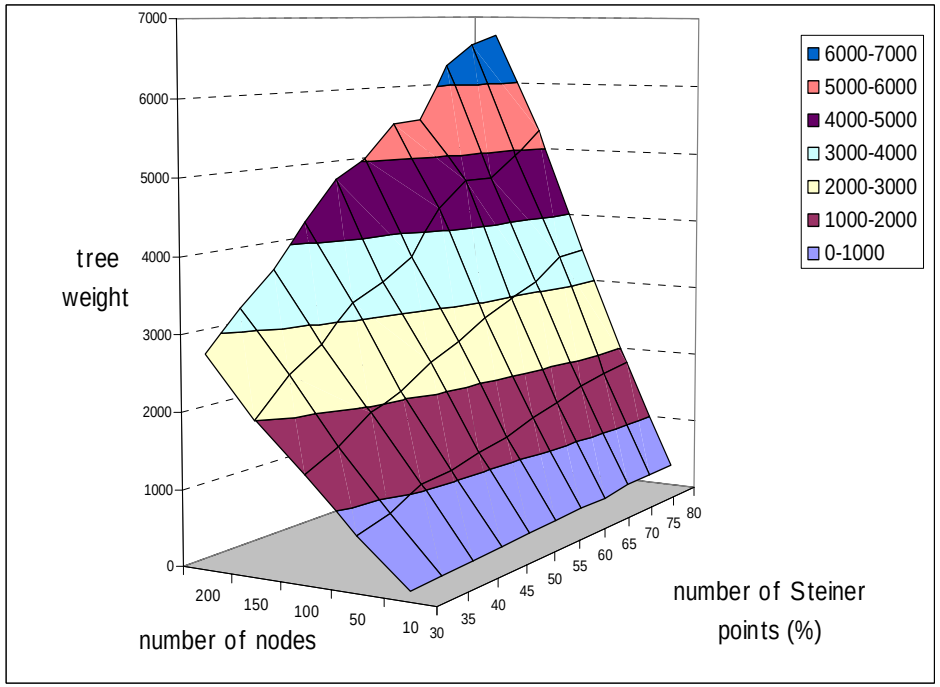
Effect of Optimisation Criterion (Cost Function)

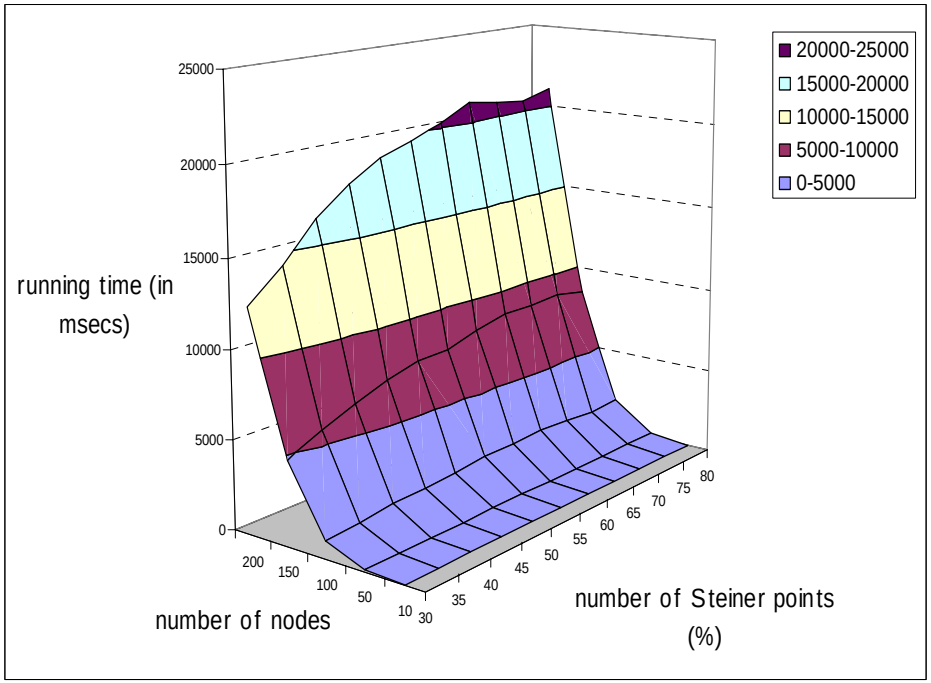




minimise $\sum_{l \in E} (F_l)^n$

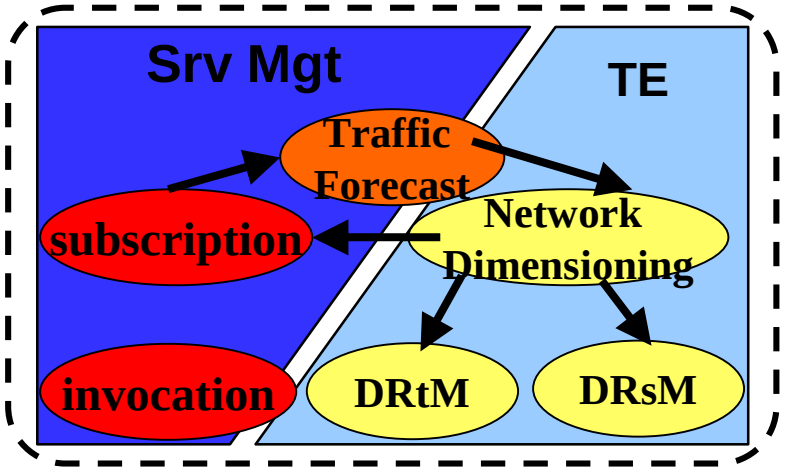
Steiner-tree Algorithm Evaluation





Since this experimentation, we have improved the above results by approximately 50% (tree weight/cost) and 20% (running time)

Conclusions



- **Service driven traffic-engineering**
- **Two-level TE approach**
  - **long-term --> guidelines for network operation**
  - **short-term --> handling traffic fluctuations**
- **Policy-driven TE operation**
  - **graceful evolution to requirements changes**
- **Interim results prove the feasibility of the approach**

TE Functionality

- **Network Dimensioning**
  - **Input: network topology, traffic matrix, policies, alarms**
  - **Objective: optimisation problem**
    - **Maintain low link cost while satisfying QoS objectives**
  - **Output in the form of configuration directives:**
    - **Explicitly routed paths (MPLS-based TE) – via DRtM**
    - **Values for the link cost metrics (IP-based TE) – directly configuring routers through combined DRtM and ND (IPTEM)**
    - **Configuration of PHB partitioning per link – via DRsM**
- **Dynamic Route Management (DRtM)**
  - **Multi-path load distribution at ingress nodes for load balancing**
  - **Explicit component in MPLS, implicit in OSPF/ECMP**
- **Dynamic Resource Management (DRsM)**
  - **Re-configuration of PHBs within allowed bounds i.e. dynamic link (re-)partitioning among service classes**
  - **Same functionality in both MPLS and IP-based TE**

Policy Information Model/Language

- **Information Model based on PCIM and PCIMe and corresponding LDAP schema**
- **High-level policy language syntax:**

  [Policy ID] [Group ID] [time period condition] [if **{condition [and] [or]}]** then **{action [and]}**
- **Example :**

  PolicyRule1: From **Time**=**0800** to **Time=1800** If **(OA == EF)** then **allocateNwBw > 30%**

  Description: At least 30% of Network Bw should be available for EF traffic between 08:00 and 18:00"

Dynamic Route Management (MPLS) – static part

- **Explicit component only in MPLS-base TE**
  - **One instance in every *edge* Label Switch Router (LSR), uses monitoring facilities within that LSR**
- **Static part gets as input from ND TT trees (and paths) with logically associated bandwidth**
- **Creates LSPs to support paths and trees**
  - **For trees, one LSP per egress leaf is created (unique path)**
  - **LSPs are created via LDP from the ingress nose with the full explicit path**
  - **No bandwidth is associated to LSPs but the edge LSR knows the logically associated bandwidth with each LSP for a path or set of LSPs realising a tree**
- **Maps "address spaces" from collected prior statistics (and SLS information) to LSPs according to their bandwidth**
  - **Configures LSR forwarding tables according to this assignment**

Dynamic Route Management (MPLS) – dynamic part

- Re-maps address spaces to alternative LSPs for load balancing according to traffic fluctuations
  - Difficult issue since existing micro-flows should be left to terminate
- Sends TT over-utilisation alarms to ND
- Sends alarms to SLS Invocation that traffic capacity to particular egress nodes is saturated
  - These act as “Congestion Notifications” to admission control
- Learns about critical PHB QoS performance at nodes crossed by its LSPs in order to take *proactive* measures
- Learns about critical end-to-end LSP QoS performance in order to take *reactive* measures
  - Knowledge about PHB and LSP performance is obtained from monitoring
  - Critical performance means “persevering conditions” and not instantaneous inability to deliver the specified QoS

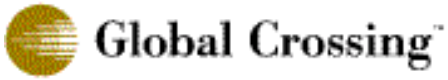
Dynamic Resource Management

- **Common in both MPLS and IP-based TE**
  - **one instance in every router, uses monitoring facilities within that router**
- **Static part configures PHB parameters**
  - **queue type – Drop tail, RED, etc.**
  - **queue parameters – buffer size, precedence/drop levels**
  - **scheduling parameters – RR, WRR, PRI, etc.**
- **Also configures performance targets set by ND per PHB**
  - **bandwidth, max loss probability, max delay**
  - **allowed bounds for dynamic (state-dependent) operation**
- **Dynamic part manages PHB resources in real-time**
  - **re-distributes (within allowed bounds) bandwidth, buffer and scheduling resources to meet dynamic traffic fluctuations**
  - **sends over-utilisation alarms to ND**

QoS Aware Monitoring & Measurement

Steven Van den Berghe

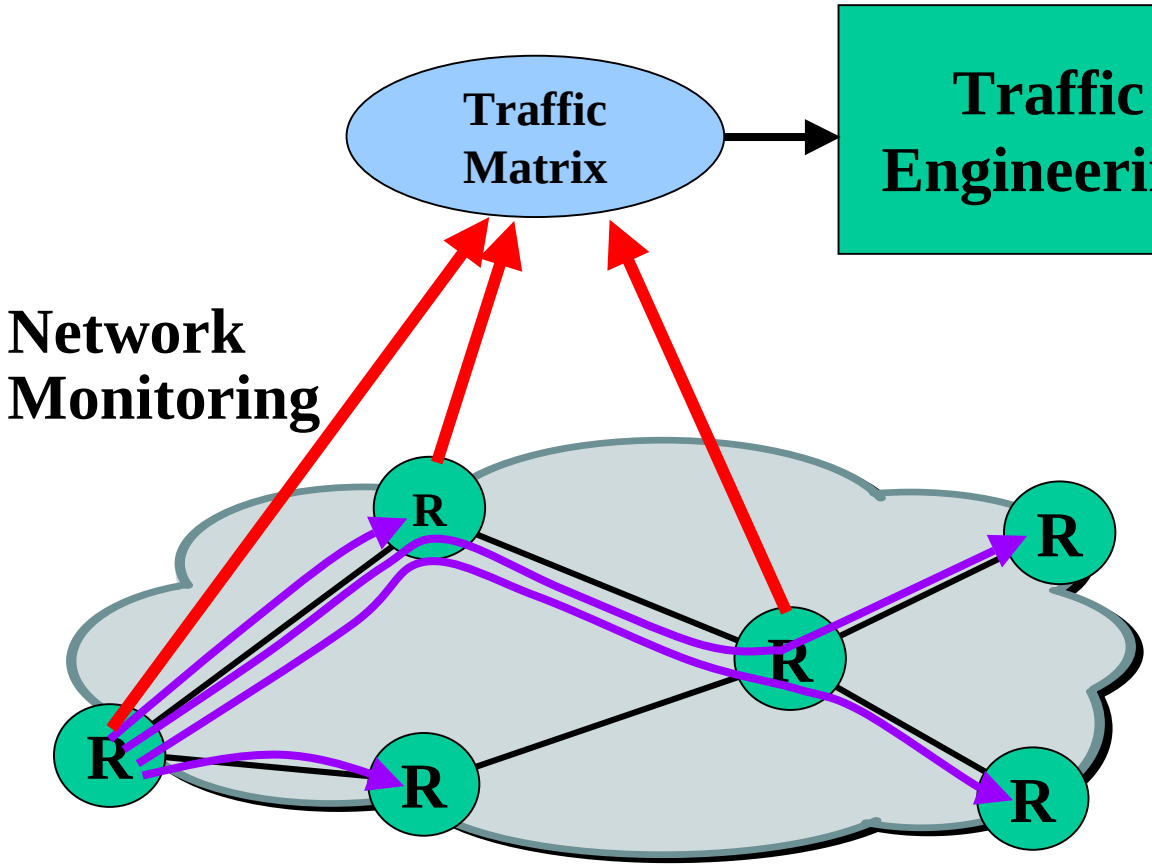
Richard Egan
Hamid Asgari



Presentation Outline

- **Role of Monitoring**
- **Architecture**
- **Design**
- **Results**
- **Scalability Features**
- **Outcomes**
- **Conclusions**

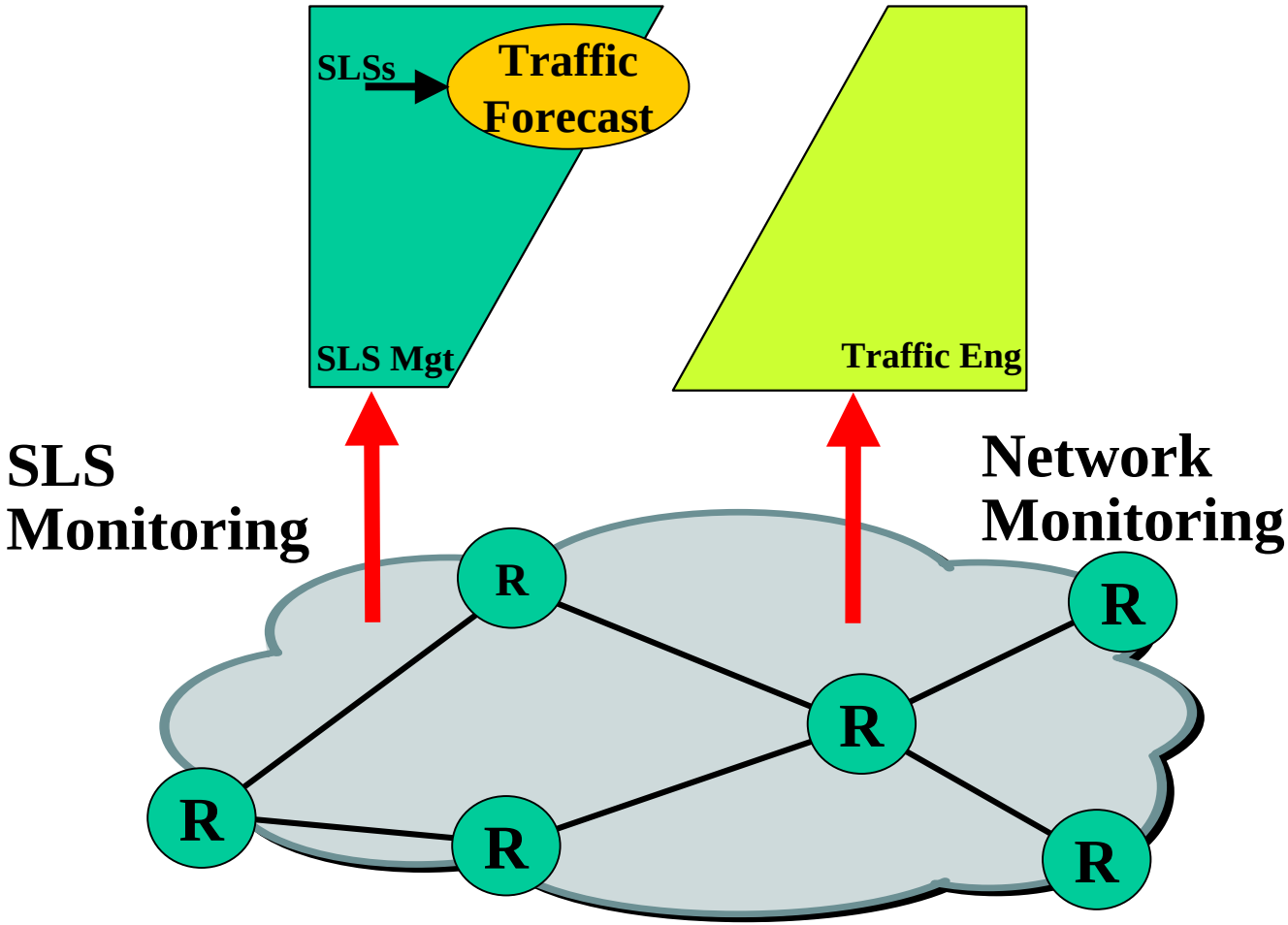
Role of Monitoring - Current



TE Methodology

- **establish LSPs**
- **monitor LSP usage**
- **create Traffic Matrix**
- **establish LSPs with BW constraint**
- **update TM via Monitoring**

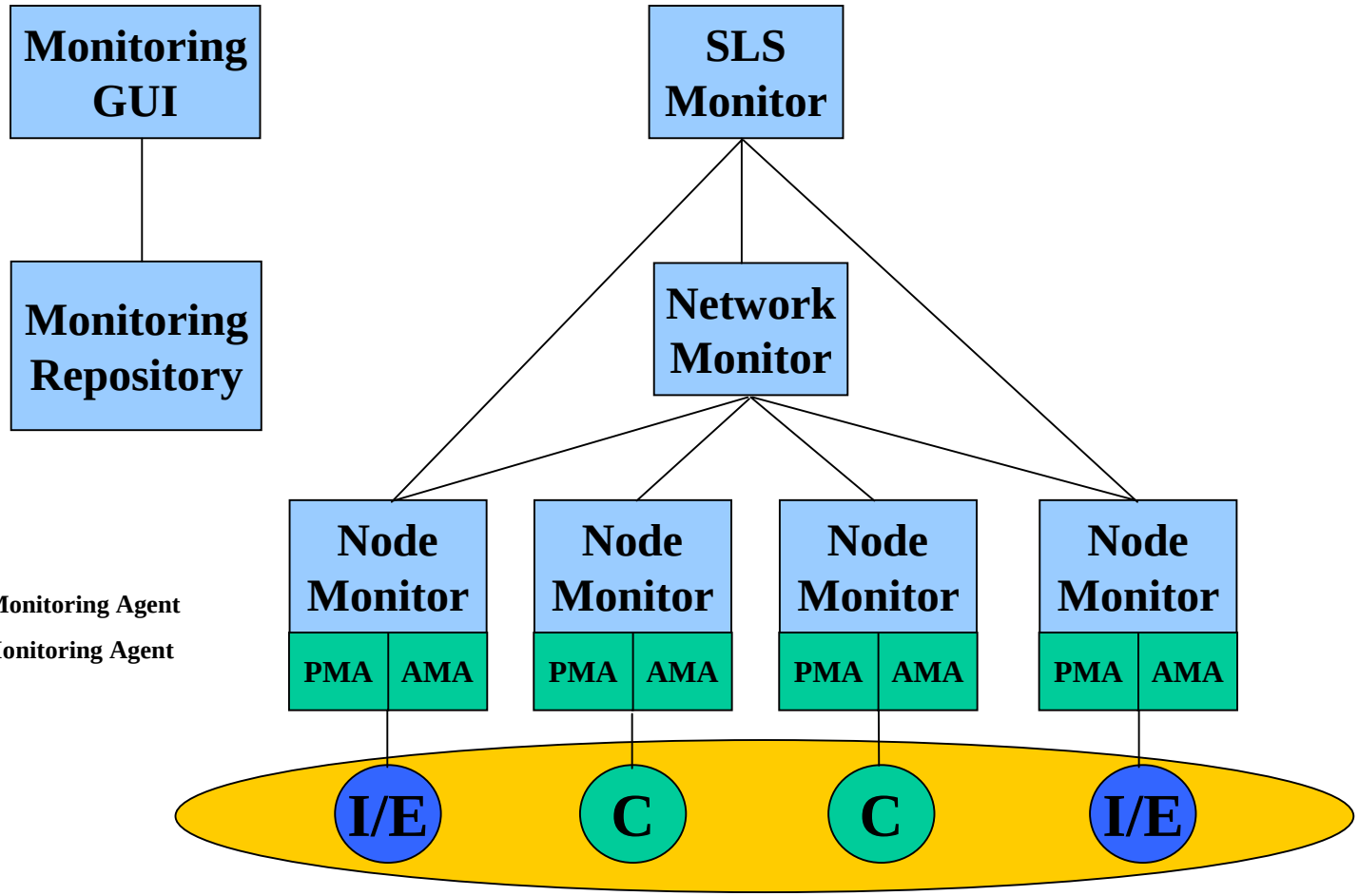
Role of Monitoring - TEQUILA



Network Monitoring

- To assist Dynamic TE to adapt to:
  - Congestion / under-utilisation
- Two primary components
  - Node Monitor
    - contains the active/passive measurement agents
    - performs all edge2edge measurements
  - Network Monitor
    - builds a physical & logical network view
    - derives path/network measurements from hop by hop results
- Relevant Metrics
  - OneWayLoss, OneWayDelay
  - PHB Bandwidth Usage, PHB Packet Discard
  - Throughput

Monitoring Architecture



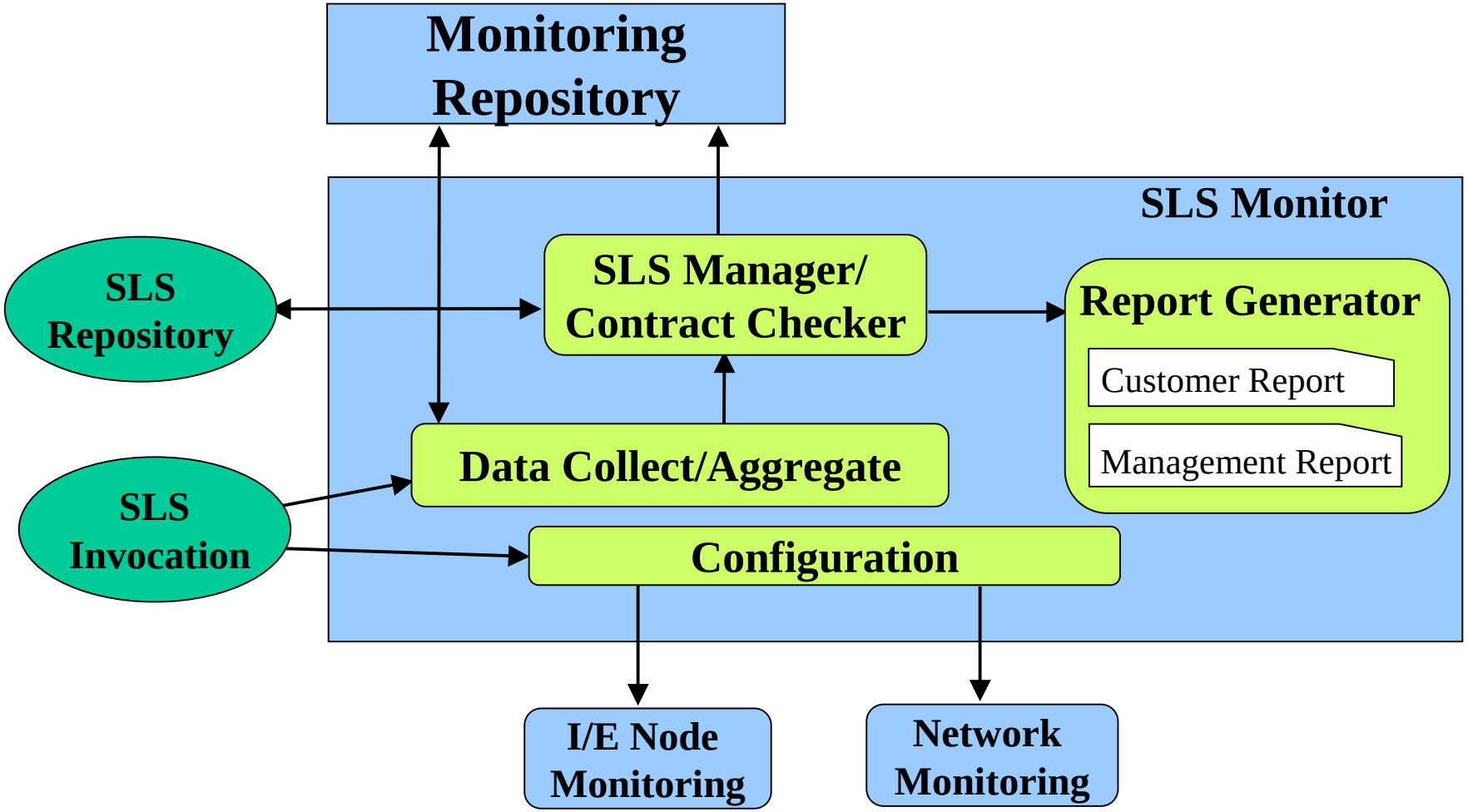
PMA Passive Monitoring Agent

AMA Active Monitoring Agent

SLS Monitoring

- In-service verification of customer services
- Provides SLS usage information to Traffic Forecasting
- SLS Monitoring is a client of Network Monitoring
- SLS Monitoring is a centralised Component
- Relevant metrics
  - OneWayLoss
  - OneWayDelay
  - Throughput
  - Offered Load

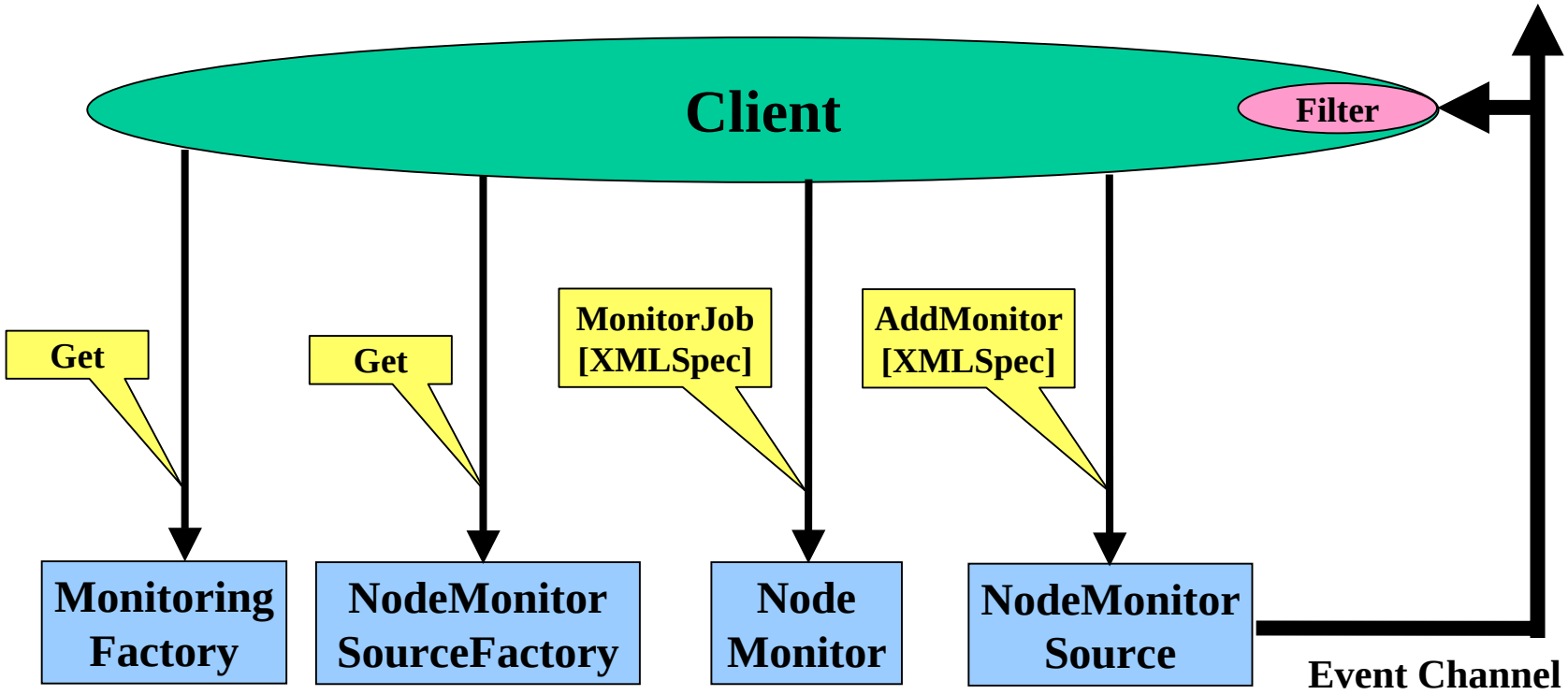
SLS Monitoring



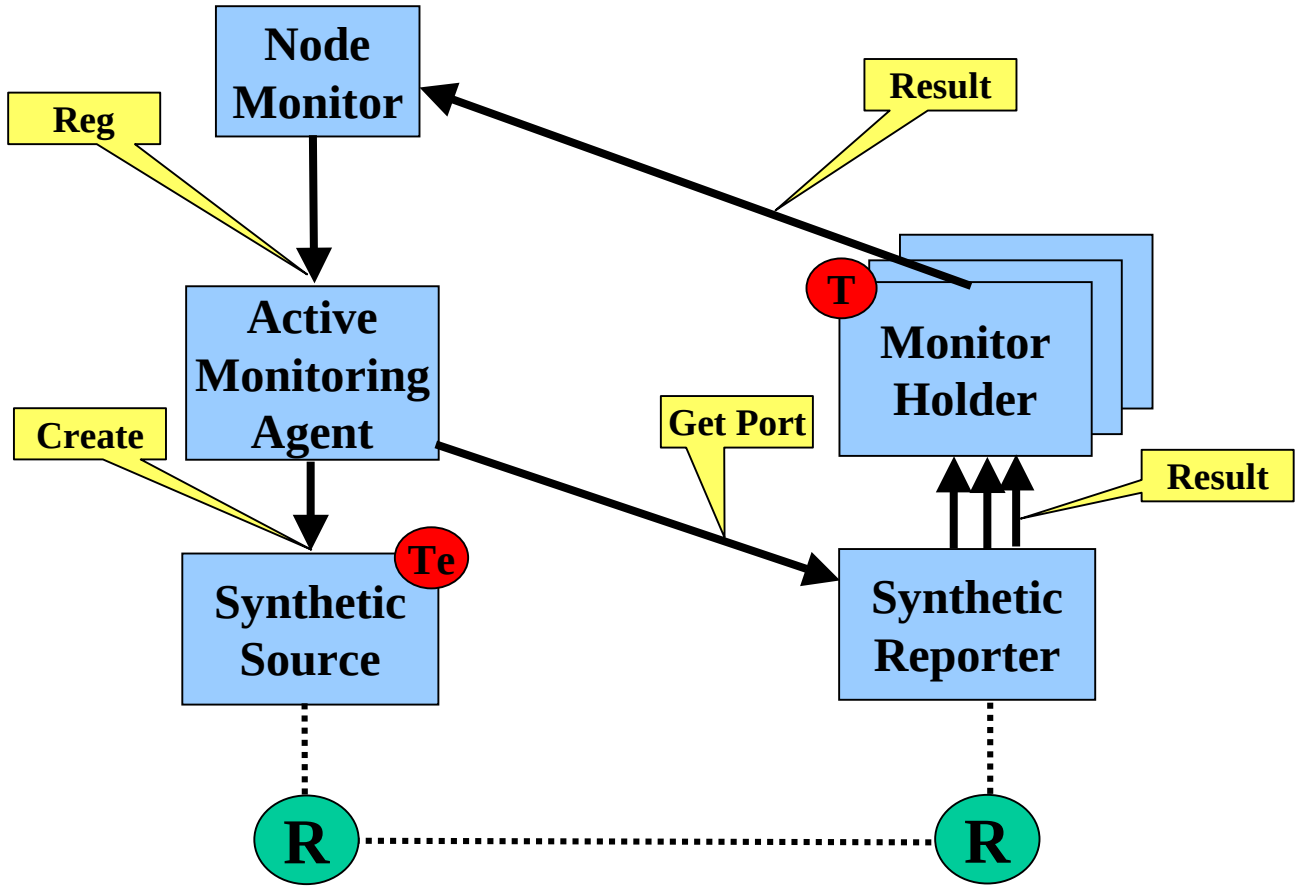
Design

- **Common Interface for Node & Network Monitoring**
  - **Register**
  - **Configure**
  - **Execute**
  - **Report Results**
- **A Monitor is created to measure a particular metric (e.g. throughput at a particular node).**
- **A MonitorJob is created to specify conditions (e.g. threshold crossings) under which notification events are generated.**

Creating a Node Monitor



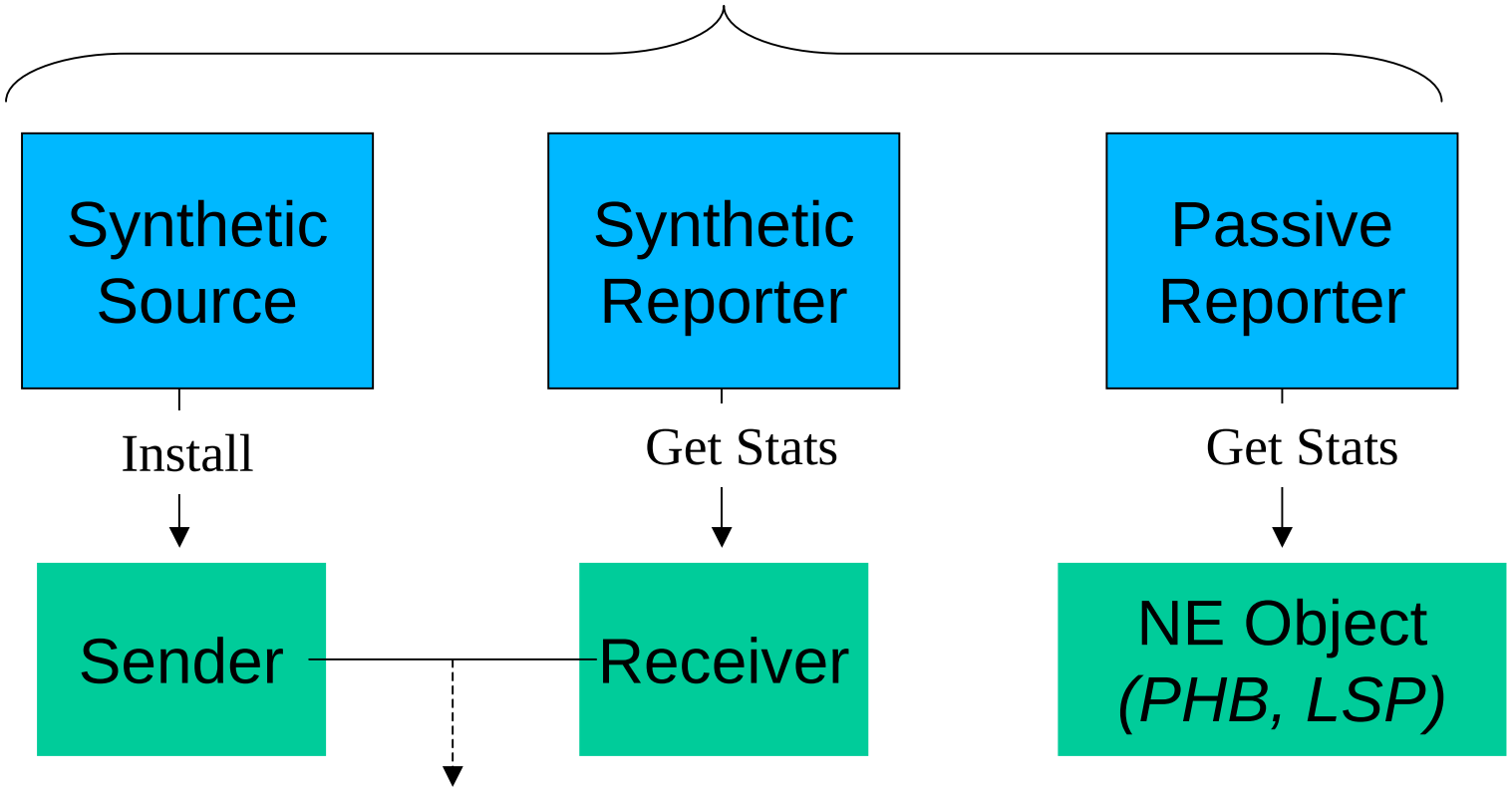
Creating an Active Monitor



Configuring Node Monitoring

- **State-of-the-art:**
  - **Passive Monitoring configured through SNMP, (emerging) COPS feedback reports, proprietary polling**
  - **Active Monitoring**
    - **Through signaling (One Way Delay Protocol)**
    - **“Top-Down”: through SNMP (emerging from IETF RMON wg)**
- **Tequila Approach: non-signaling**
  - **Same configuration methodology for active and passive**
  - **Extra features in OWDP-protocol (e.g. inter-domain related) are no requirement for intra-domain**
  - **Same configuration methodology as other low-level Tequila components (e.g. tunnel establishment and PHB configuration)**

Basic Entities

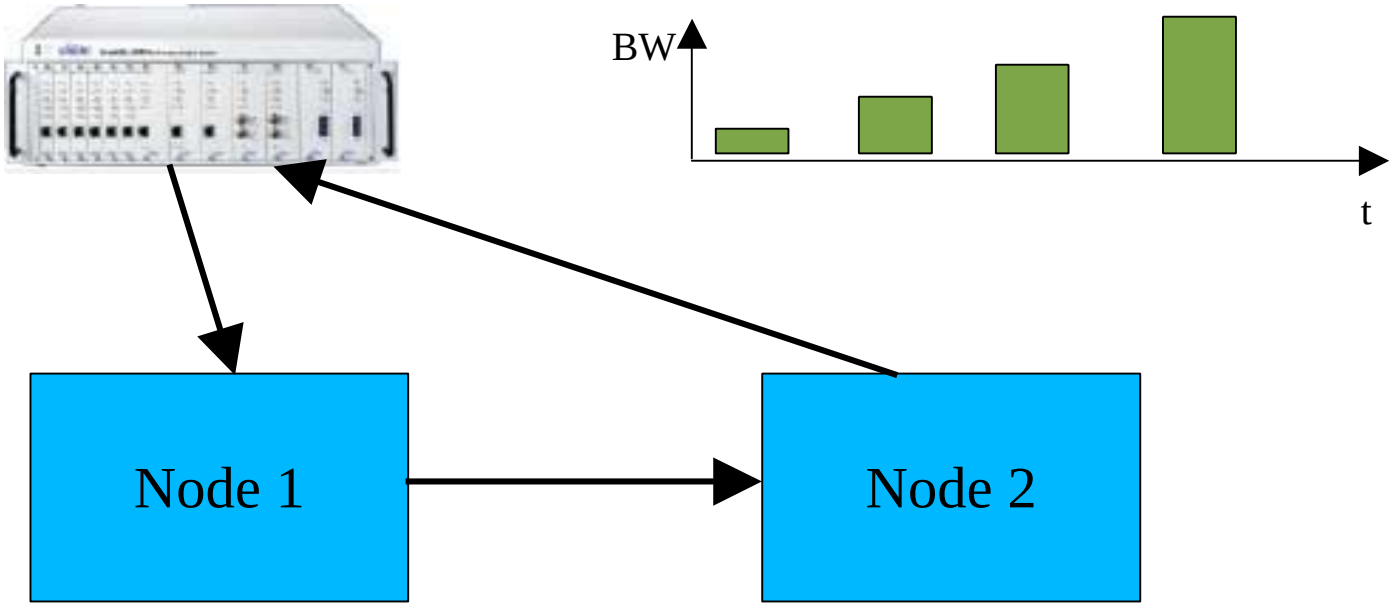


Connection determined by IP-addresses and Destination UDP address

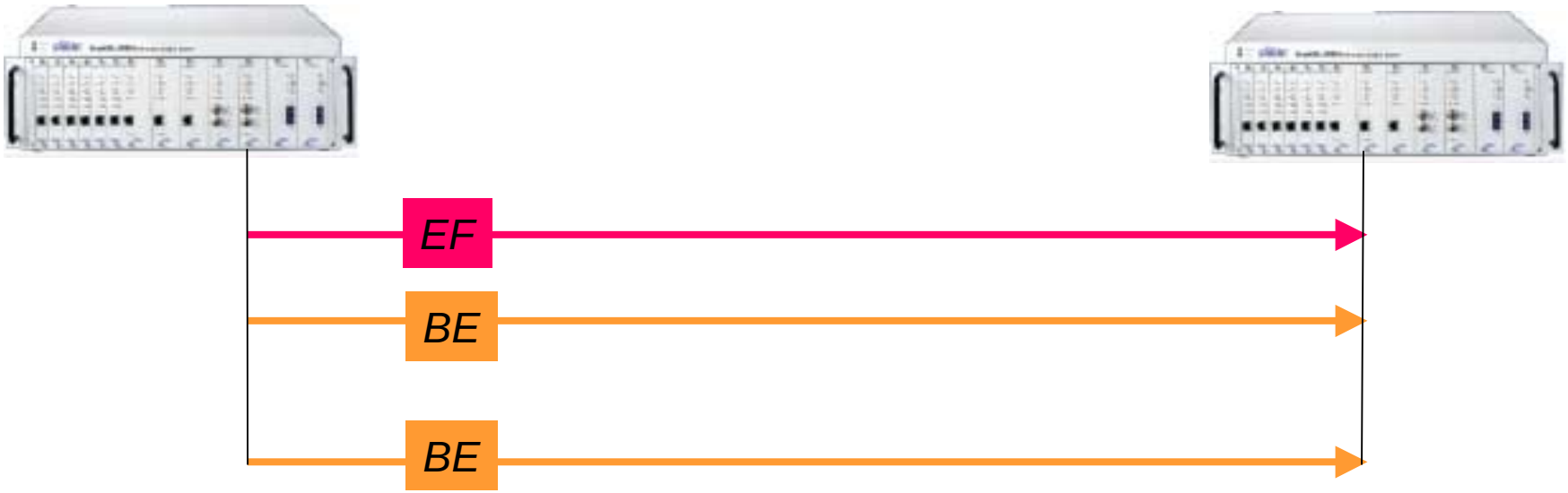
Test Set-up

SmartBits:

- Generates traffic between 1 and 50 Mb/s (step 10Mb/s)
  - no congestion, just functional observation
- Test time: 60s per step

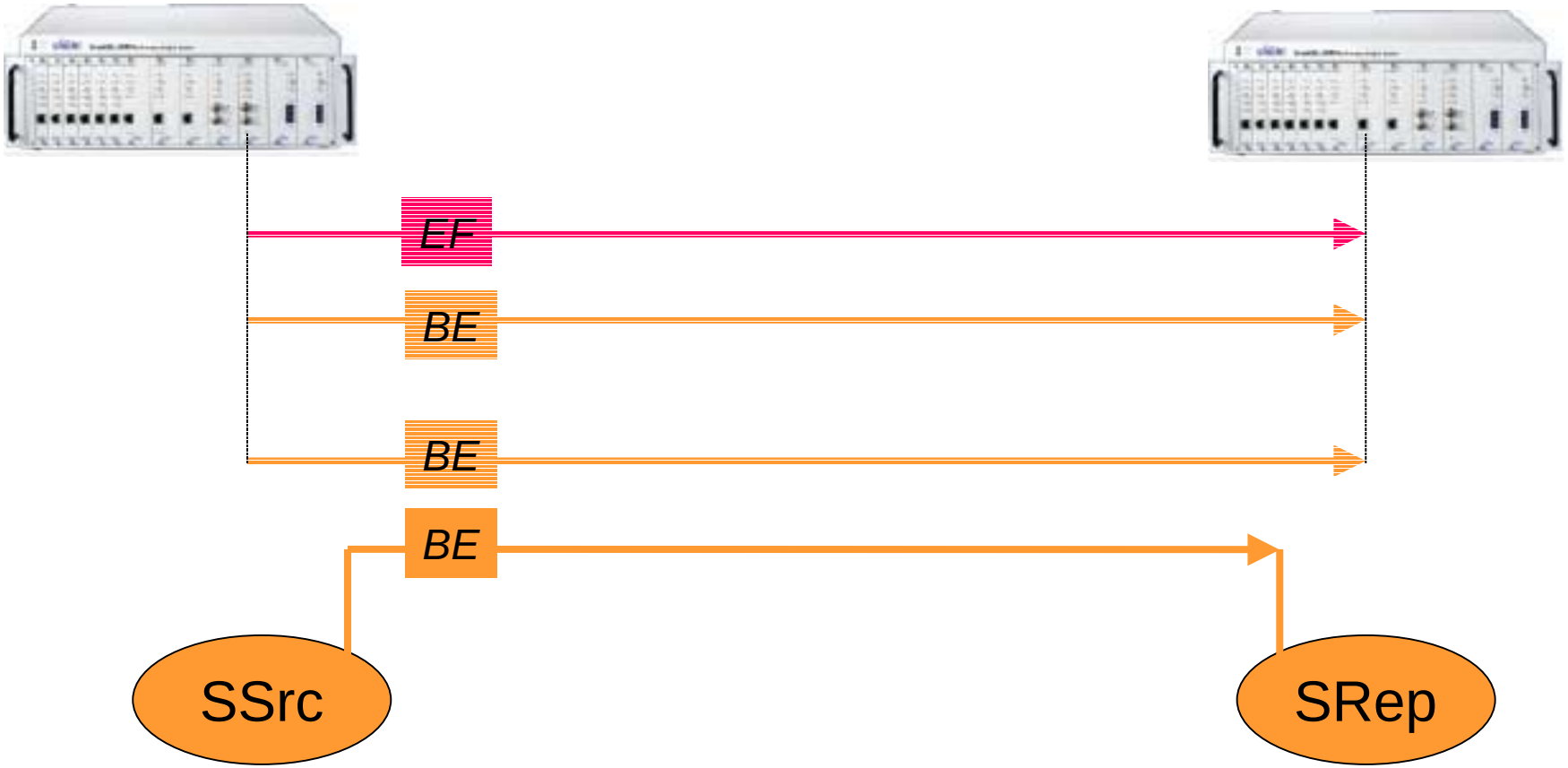


Traffic Set-up



Divided into 3 Streams

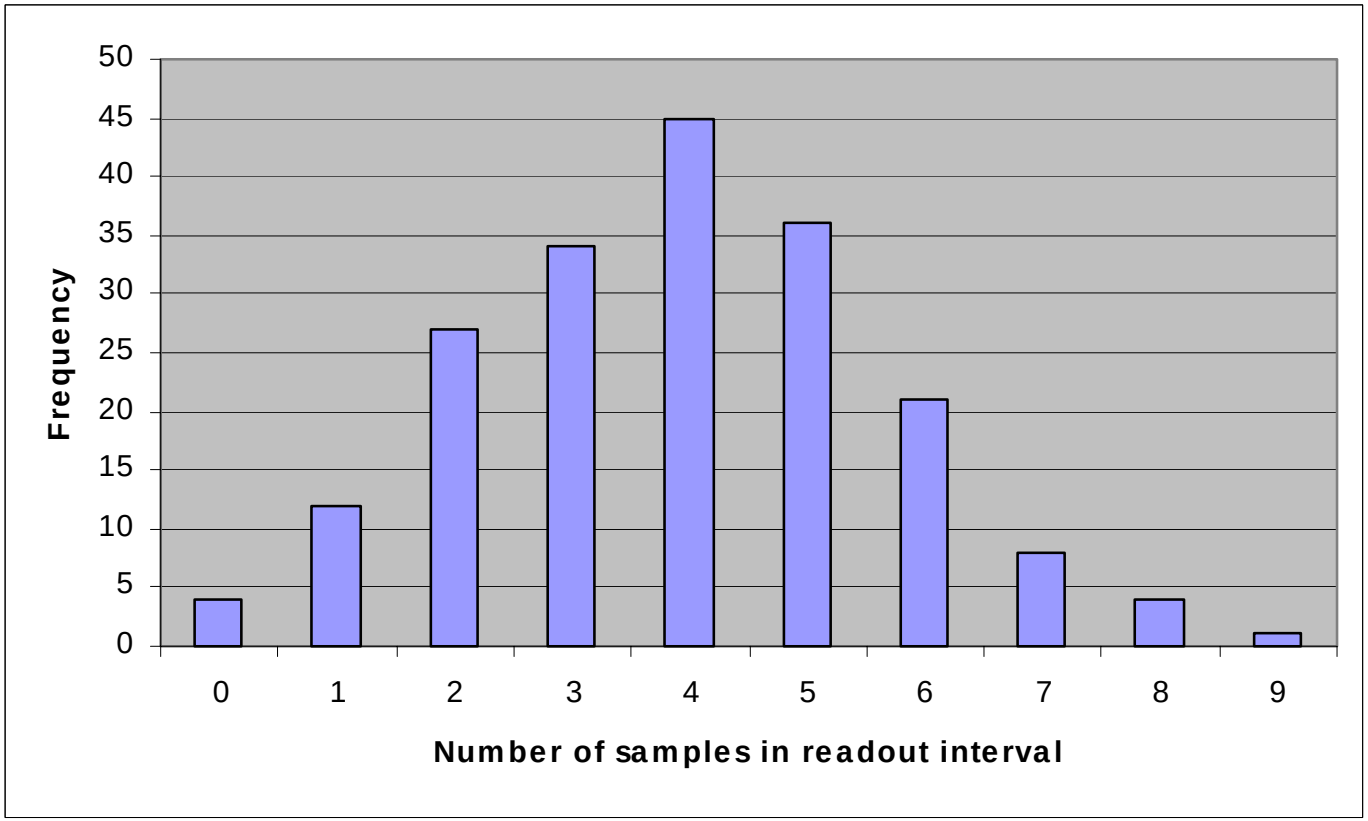
Active Measurement



2 packets/sec

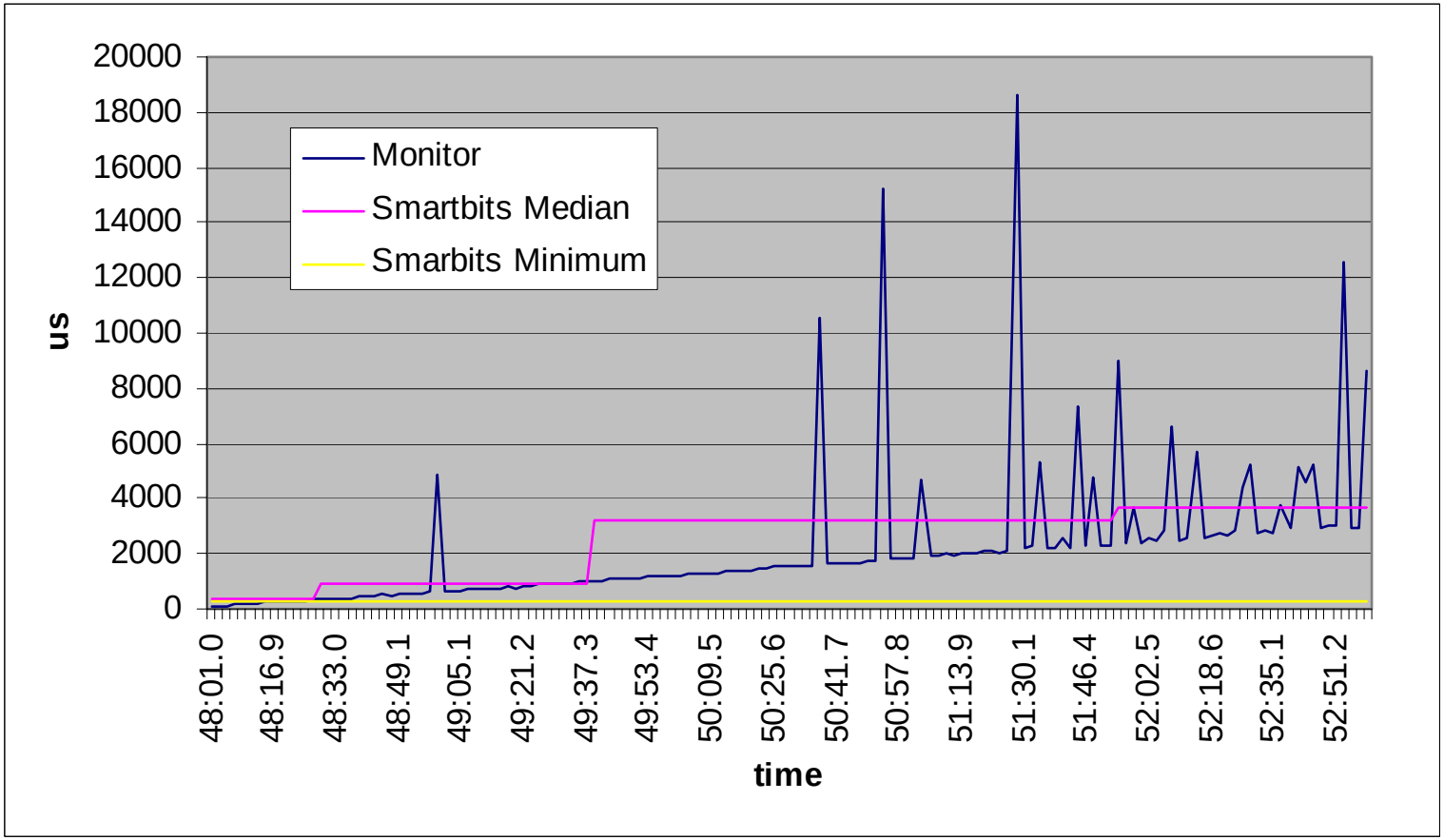
Average Delay over 2 secs

Active Measurement

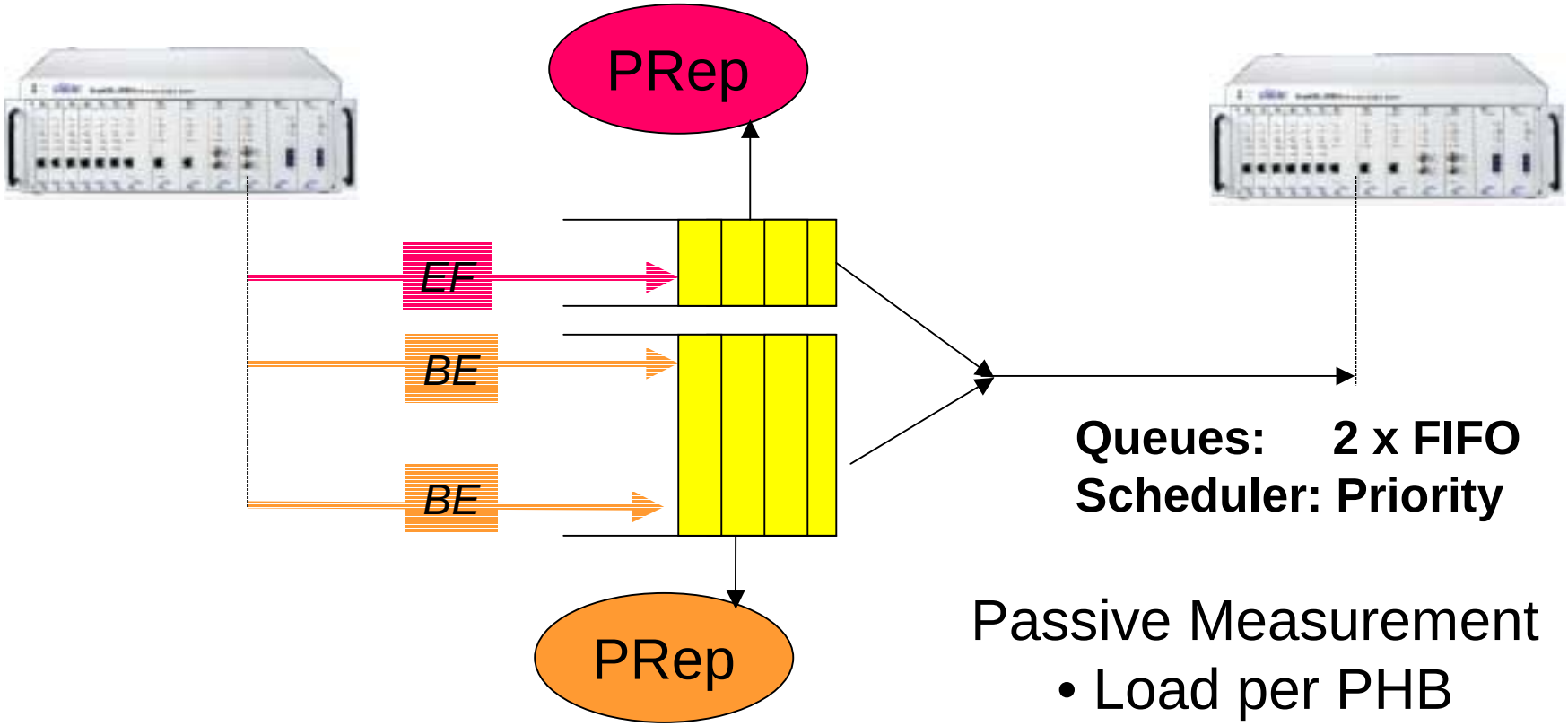


Exponential Inter-arrival

Active Measurement



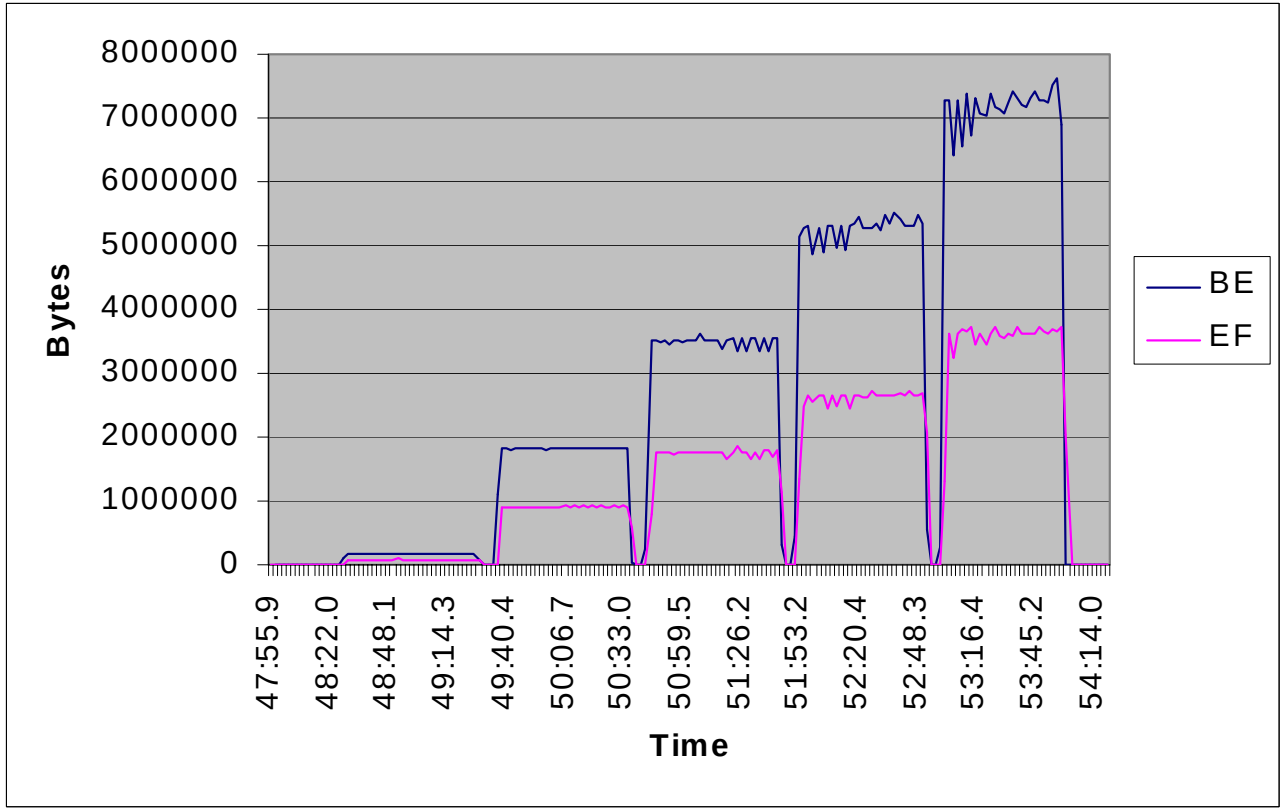
Passive Measurement



Queues: 2 x FIFO
Scheduler: Priority

Passive Measurement
- Load per PHB
- Readout = 2 sec

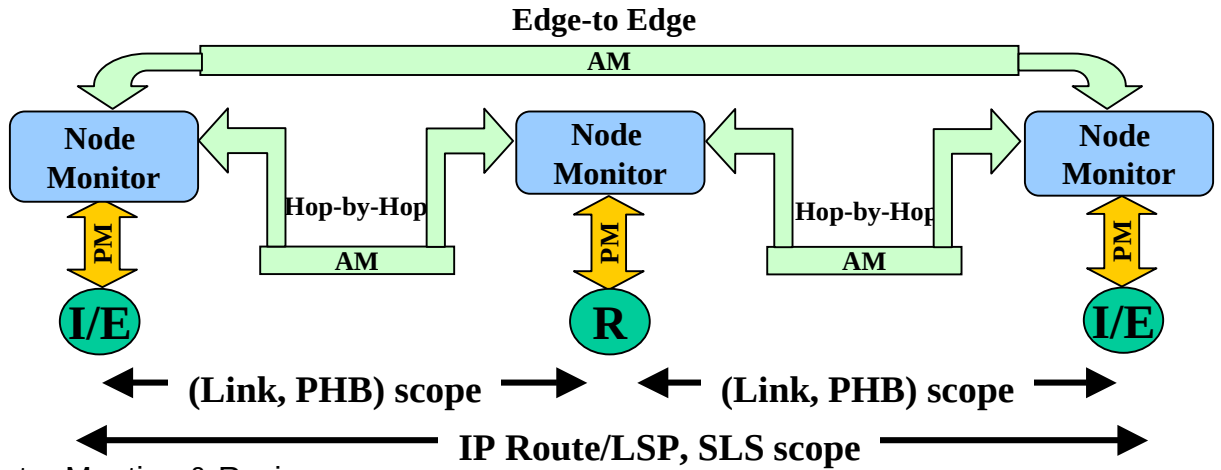
Passive Measurement



Final Ratio **BE : 66.63%**
EF: 33.37%

Scalability (1)

- **SLS Monitor uses Network Aggregate Measurements**
  - **Per LSP monitoring is more scalable than per SLS**
  - **Combine with per-SLS Ingress/Egress measurements**
    - **throughput / offered load**
- **Use Hop by Hop Measurements**
  - **Reduce volume of synthetic traffic**
  - **Aggregate hop measurements to get E2E measurement**



AM: Active Measurement
PM: Passive Measurement

Scalability (2)

- **Distributed Probes**
  - **Event notification**
  - **Client interface enables shared use of probes**
  - **(near) Real time response time**

Observations

- **Intra-domain only**
  - **Sufficient for most business services (VPN, VLL…)**
  - **OWDP Control Signalling not needed**
- **Access Network not addressed**
  - **a limitation as AN adds delay and loss**
  - **out of scope for TEQUILA**
- **Node Monitor Agents not integrated in routers**
- **Burden on Ingress/Egress routers**

Publications

- **A Monitoring and Measurement Architecture for Traffic Engineered IP Networks**
  - **IST 2001, Tehran, Sept 1-3, 2001**
- **A Framework for Internet Traffic Engineering Measurement**
  - **<draft-ietf-tewg-measure-01.txt>**
  - **Official TE WG document**
  - **Last version before ‘Final Call’ for Informational RFC**

Conclusions

- **Solution for**
  - **Dynamic TE**
  - **SLS Monitoring**
- **Design Features**
  - **Common interface for Node & Network Monitoring**
  - **Common configuration approach for Active & Passive Agents**
  - **Separation of Monitor & MonitorJob**
- **Scalability Features**
  - **Flexible SLS Monitoring based on E2E or Hop by Hop**
  - **Network Monitoring based on event notification**

QoS routing over the Internet: a BGP4-based approach

Christian JACQUENET

France Telecom R & D

christian.jacquenet@rd.francetelecom.com

Dresden Nov. 21, 2001

QoS routing over the Internet

- **Agenda:**
  - **Motivation and requirements**
  - **Proposal**
  - **Simulation work and preliminary results**
  - **Issues**
  - **Ongoing work**
  - **Conclusion**

Motivation and requirements

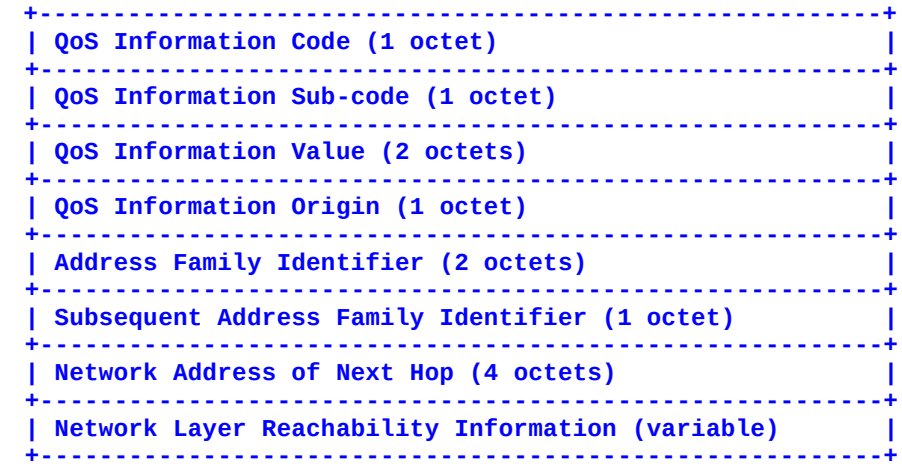
- QoS policy enforcement is currently restricted to the scope of an AS
- QoS information needs to be exchanged between domains
  - Existing BGP4 attributes can help in providing some kind of « QoS indication »
    - *E.g.* the « PREFER_ME » and « AVOID_ME » global values of the COMMUNITIES attribute
  - But a finer granularity would be useful
- Allow for a smooth migration
  - Gradual deployment of QoS route computation over the Internet
- Keep the approach scalable

Proposal

- **Use the BGP4 protocol for conveying QoS-related information between domains to:**
  - **Enable QoS-based route selection processes**
  - **Enhance peering agreements for the deployment of value-added IP services across domains**
  - **Contribute to the enforcement of end-to-end QoS policies**
- **Introduce a new optional transitive attribute:**
  - **The QOS_NLRI attribute**

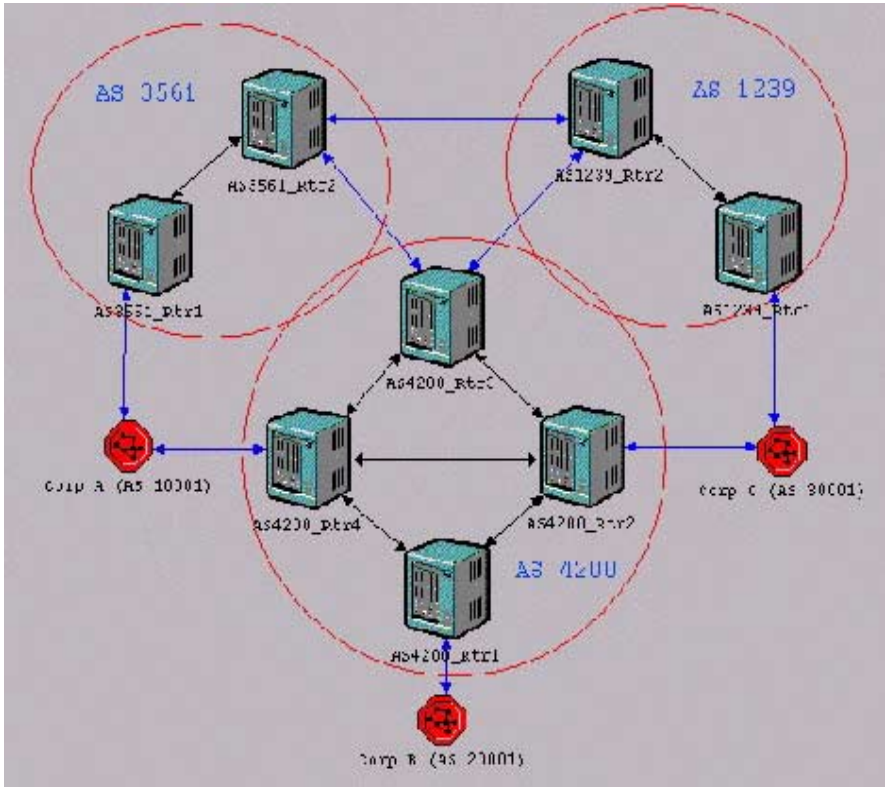
The QOS_NLRI attribute

- **Advertise « QoS routes », *i.e.* routes that can be depicted with specific QoS information**
  - ***E.g.* “route to network N1 experiences a 100 ms one-way transit delay”**
- **Provide QoS information associated to the destination prefixes**
  - ***E.g.* “EF-marked datagrams may use this route to network N2”**



Network modeling

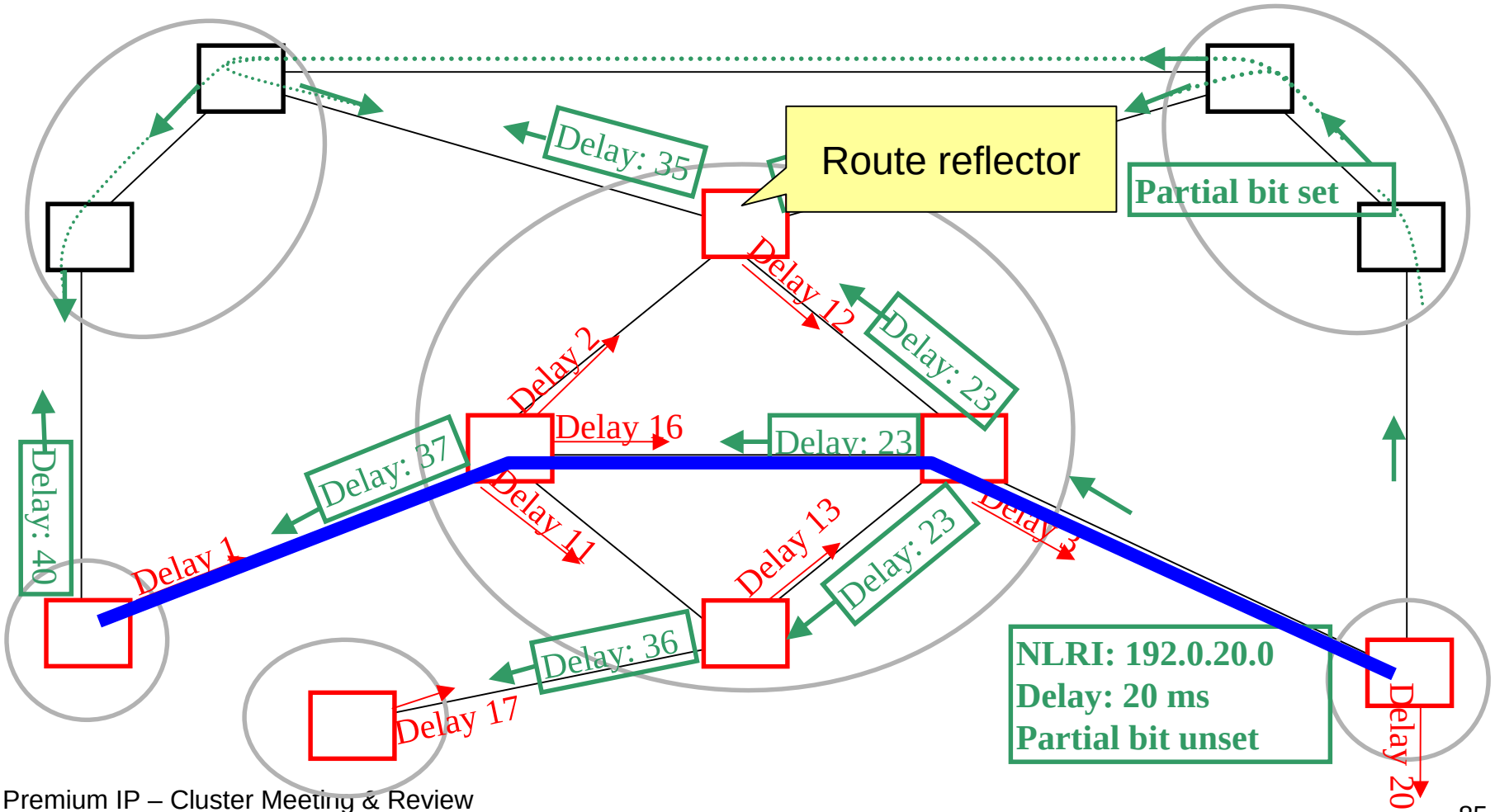
- **The network model:**
  - **A mixed environment**
    - **QOS_NLRI enabled and “classical” BGP peers**
    - **CPE/PE/P node taxonomy**
  - **Multiple domains**



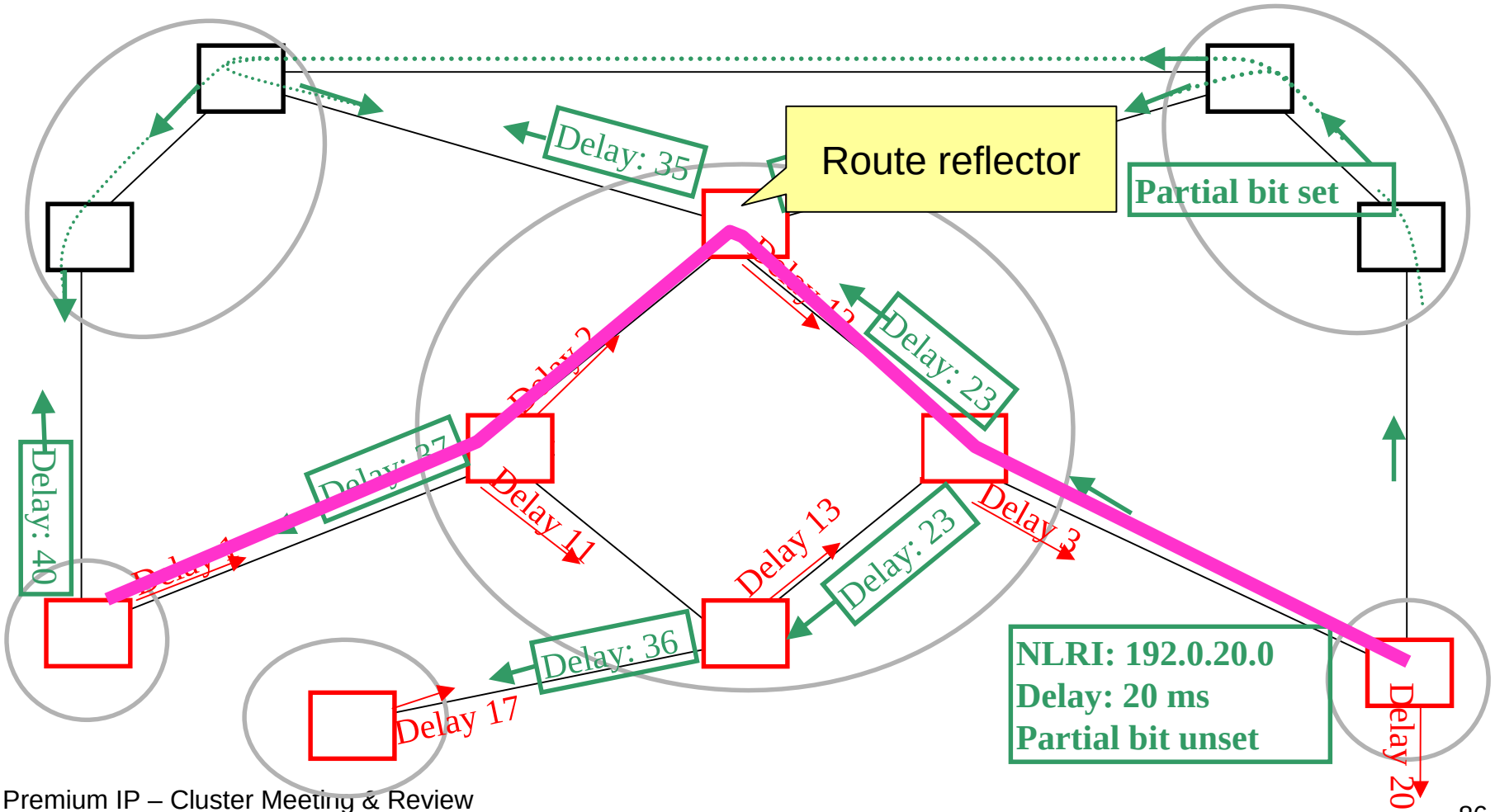
BGP peer

Access node

Example of route selection

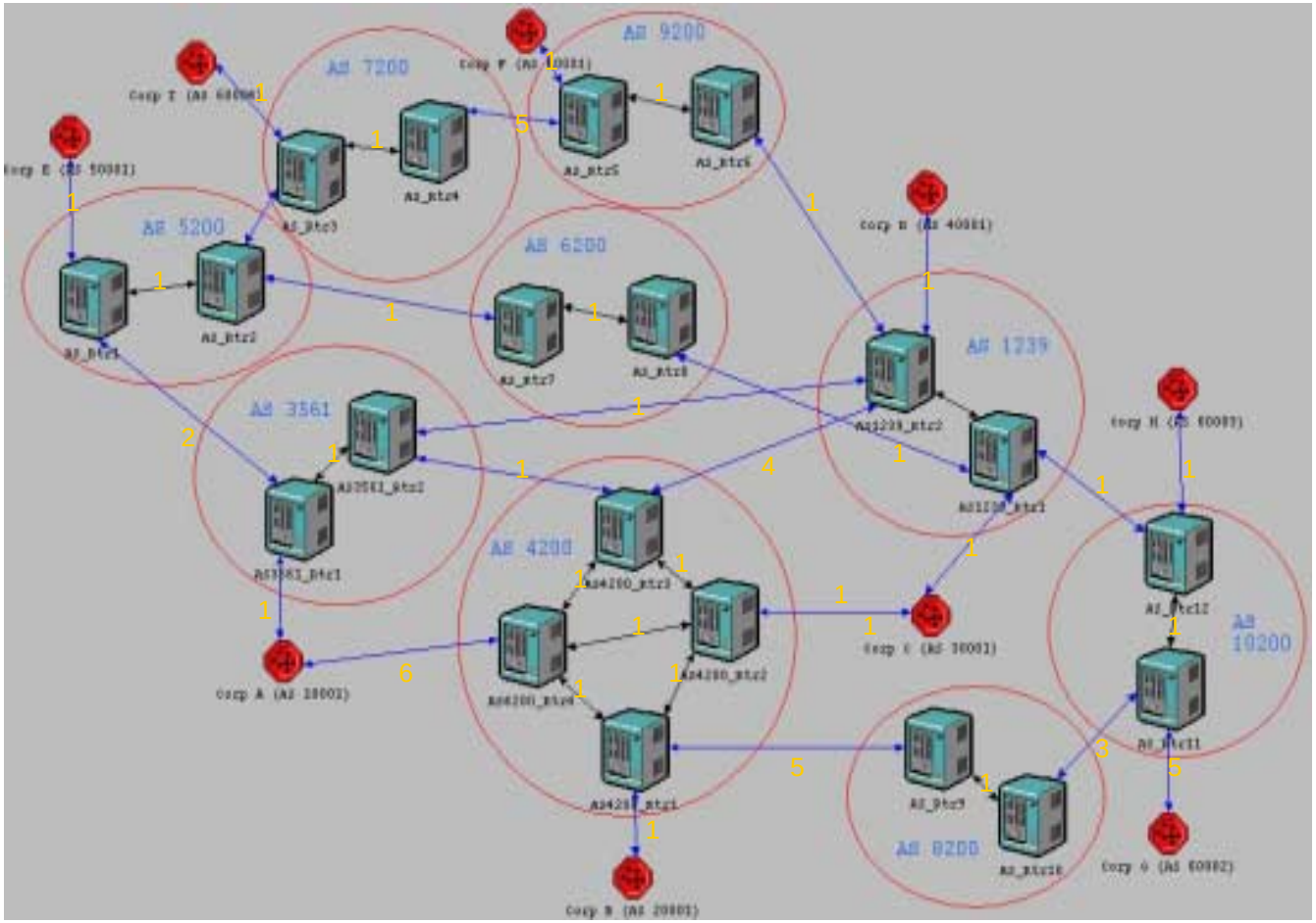


Example of route selection



192.0.20.0

Simulated network

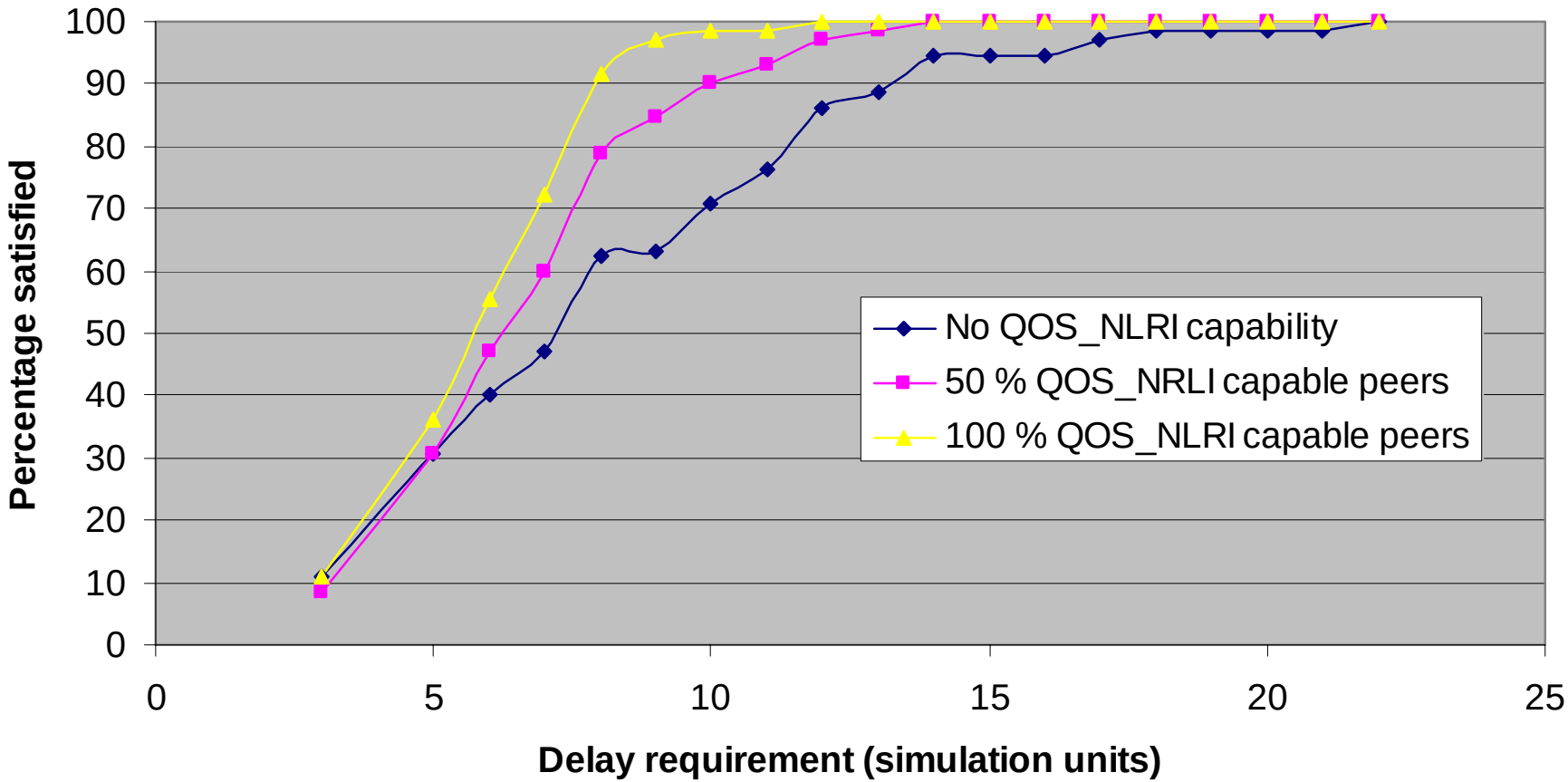


Simulation parameters

- **Percentage of BGP speakers that are QOS_NLRI capable**
  - **0%: reference network (as a collection of autonomous systems)**
  - **0%<x%<100%: x% of the BGP peers are QOS_NLRI-enabled**
- **Delay requirements for traffic between a source and a destination**
  - **Strongest (lowest-delay) requirements have less chance to be satisfied**
- **Transit delays on the links**
  - **The higher the delay on the links, the lower the percentage of serviced SLSs**

Preliminary simulation results

- **Satisfying delay requirements:**



Issues

- **Scalability:**
  - **How frequently should UPDATE messages be sent, according to changes of the « bandwidth conditions »?**
  - **Aggregation capabilities:**
    - **How to provide QoS indication with aggregated routes, and what would be the aggregation criteria?**
- **Stability:**
  - **PHB Id. and delay-related information should not yield flapping conditions**
    - **Dynamics of bandwidth information remains an issue**
- **Confidentiality of QoS information:**
  - **Already made publicly available by most ISPs**
    - **By means of looking glasses, for example**

Ongoing work

- **Convey additional QoS information**
- **Update draft for the next IETF meeting**
  - **See current** `draft-jacquenet-qos-nlri-03.txt/pdf` **on TEQUILA web site**
- **Ongoing prototype development**
  - **Based upon Zebra's code (www.zebra.org)**
- **Additional simulation results by Q1 2002**
  - **Submit an applicability draft to the IETF**

Conclusion

- **On the approach:**
  - **NO modification of the BGP4 protocol**
- **On the simulation:**
  - **Preliminary results are encouraging**
- **On the remaining issues:**
  - **Technical feasibility has been demonstrated**
  - **Further simulation planned to investigate the scalability aspects**