UNIVERSITEIT GENT

FACULTY OF ENGINEERING

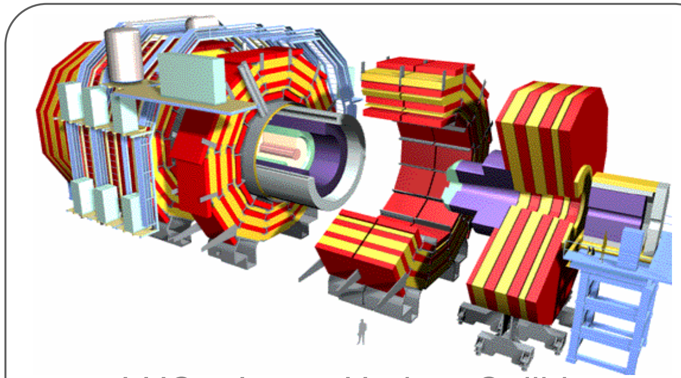
Delivering the Grid Promise with Optical Burst Switching

Chris Develder

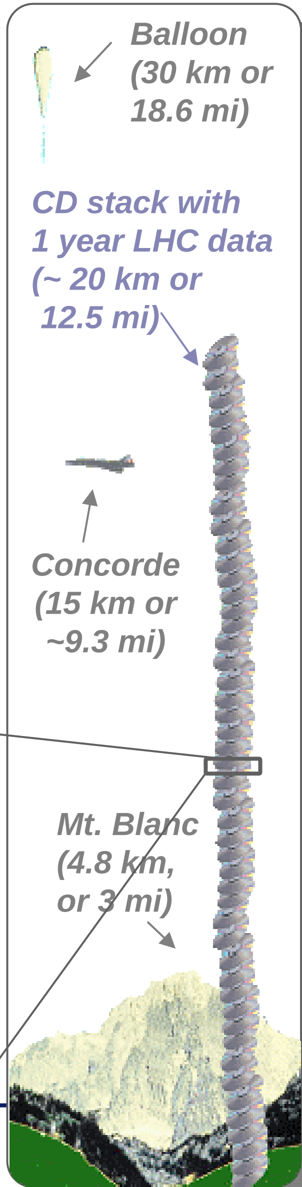
M. De Leenheer, T. Stevens, J. Baert, P. Thysebaert, F. De Turck, B. Dhoedt, P. Demeester

Introduction (1)

- **eScience:**
  - By 2015 it is estimated that **particle physicists** will require exabytes (10^{18}) of storage and petaflops per second of computation [1]
  - CERN's LHC Computing Grid (LGC) will start operating in 2007 and will generate **15 petabytes** annually (that's ~2Gbit/s) [2]



LHC = Large Hadron Collidor: particle accellerator



Introduction (2)

- **Consumer service:**
  - Eg. **video editing**: 2Mpx/frame for HDTV, suppose effect requires 10 flops/px/frame, then evaluating 10 options for 10s clip is <u>50 Gflops</u> (today's high performance PC: <5 Gflops/s) [3]

<u>Online gaming</u>: e.g. Final Fantasy XI: 1.500.000 gamers

*<u>Virtual reality</u>: rendering of 3*10^8 polygons/s → 10^4 GFlops*

<u>Multimedia editing</u>

Introduction (4)

- **Conclusion:**
  - Grid opportunities ranging from academia over corporate business to home users
  - Optical data speeds ≥ internal PC bus speeds
    ⇒ network speed no bottleneck

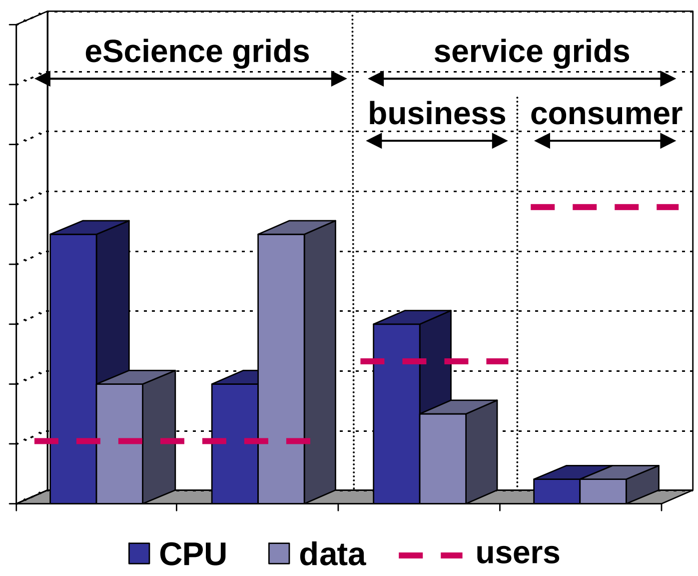
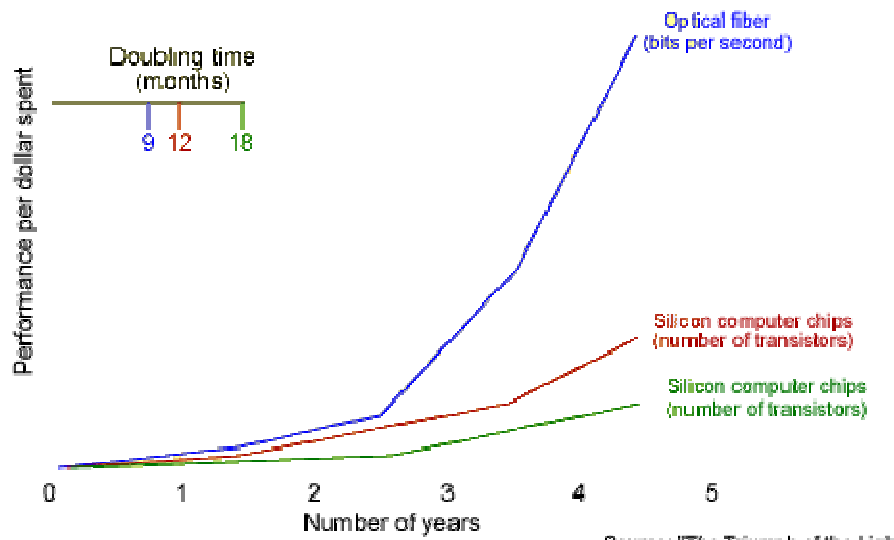
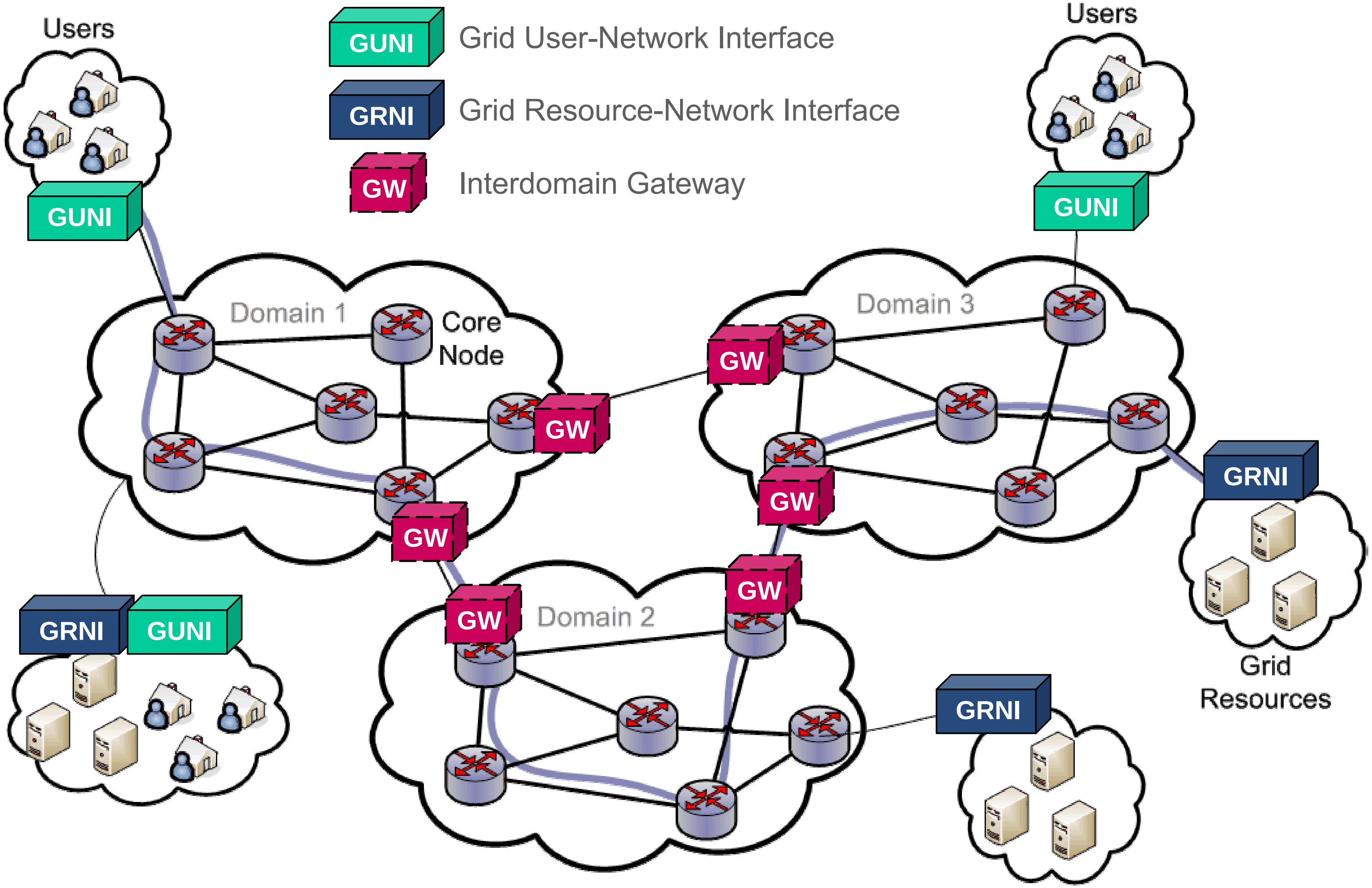


Figure taken from [5]



Source: "The Triumph of the Light," Scientific American, January 2001

Grid Network Infrastructure



■ GUNI = Grid User Network Interface

- Interoperable procedures between user and Grid
- Submits jobs (with requirements, e.g. data/CPU, time constraints, …)
- Directly via control plane, or middleware

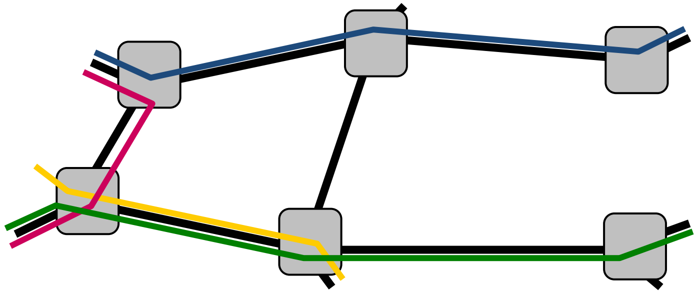
■ GRNI = Grid Resource Network Interface

- Resources can dynamically enter/leave network
- Announces processing and/or storage resources
- Signaling & control interface between NE and network

Optical Network Architecture

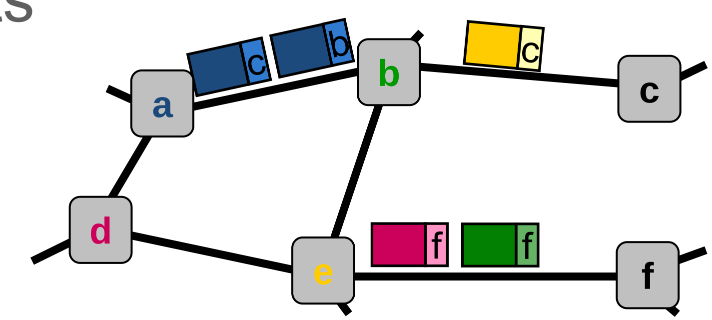
Optical Circuit Switching (OCS)

- continuous bit-stream
- pre-established light-paths
- should be dynamic



Optical Burst/Packet Switching (OBS/OPS)

- chunks of bits, in bursts/packets
- forwarding based on header
- e.g. label switching, GMPLS



Hybrids

Figures taken from [6]

Optical Circuit Switching

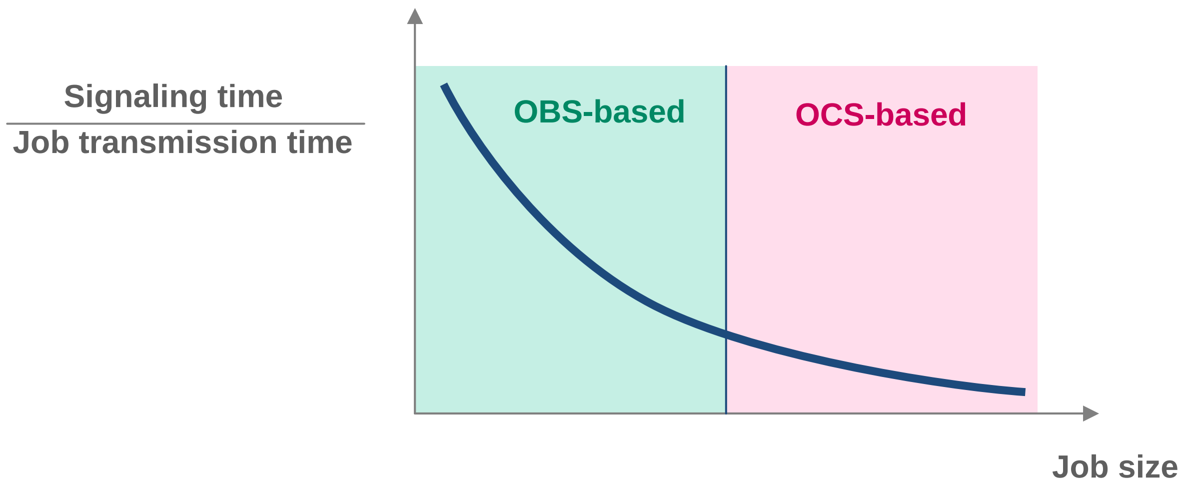
- **Pro:**
  - ✓ Guaranteed service quality once set-up (cf. reserved lambda), thus fixed latency, no jitter, etc.
  - ✓ Fixed signaling overhead, independent of (large) job size
- **Con:**
  - ✗ Signaling overhead† not acceptable for relatively small jobs
    - ✗ **Requires (complex) grooming if frequent set-up and tear-downs are to be avoided (i.e. if too slow)**
  - ✗ Less flexible, dynamic than OBS/OPS, cf. light-path set-up and tear-down

†: [7] cites 166ms/switch → RSVP-TE speedup needed

OBS/OPS

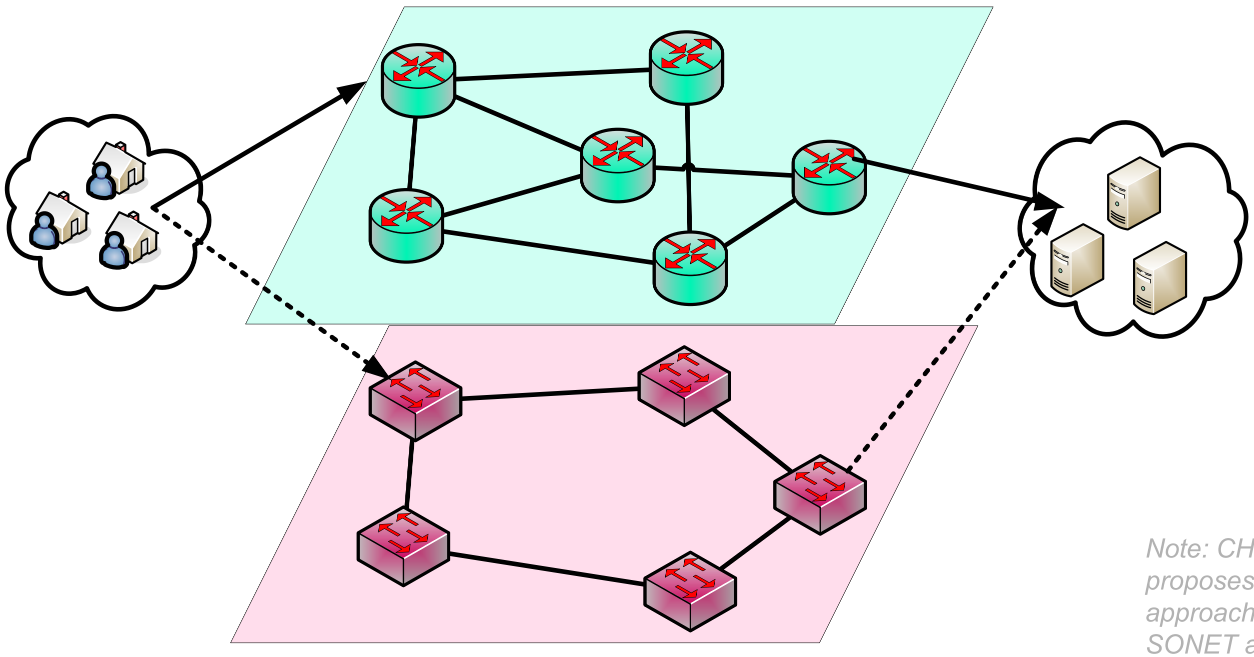
- **Pro:**
  - ✓ Extremely flexible, dynamic
  - ✓ Inherent statistical multiplexing of available bandwidth (over multiple lambdas)
- **Con:**
  - ✗ Packet/Burst header processing overhead
    - ✗ **Requires job aggregation if job size too small compared to header overhead**
  - ✗ Difficult to deliver strict QoS guarantees without 2-way reservation
  - ✗ Technology not that mature

- **Choosing between OCS and OBS depends on…**
  - Optical technology (OBS requires faster switches, burst mode Rx/Tx and regenerators, …)
  - Job sizes:



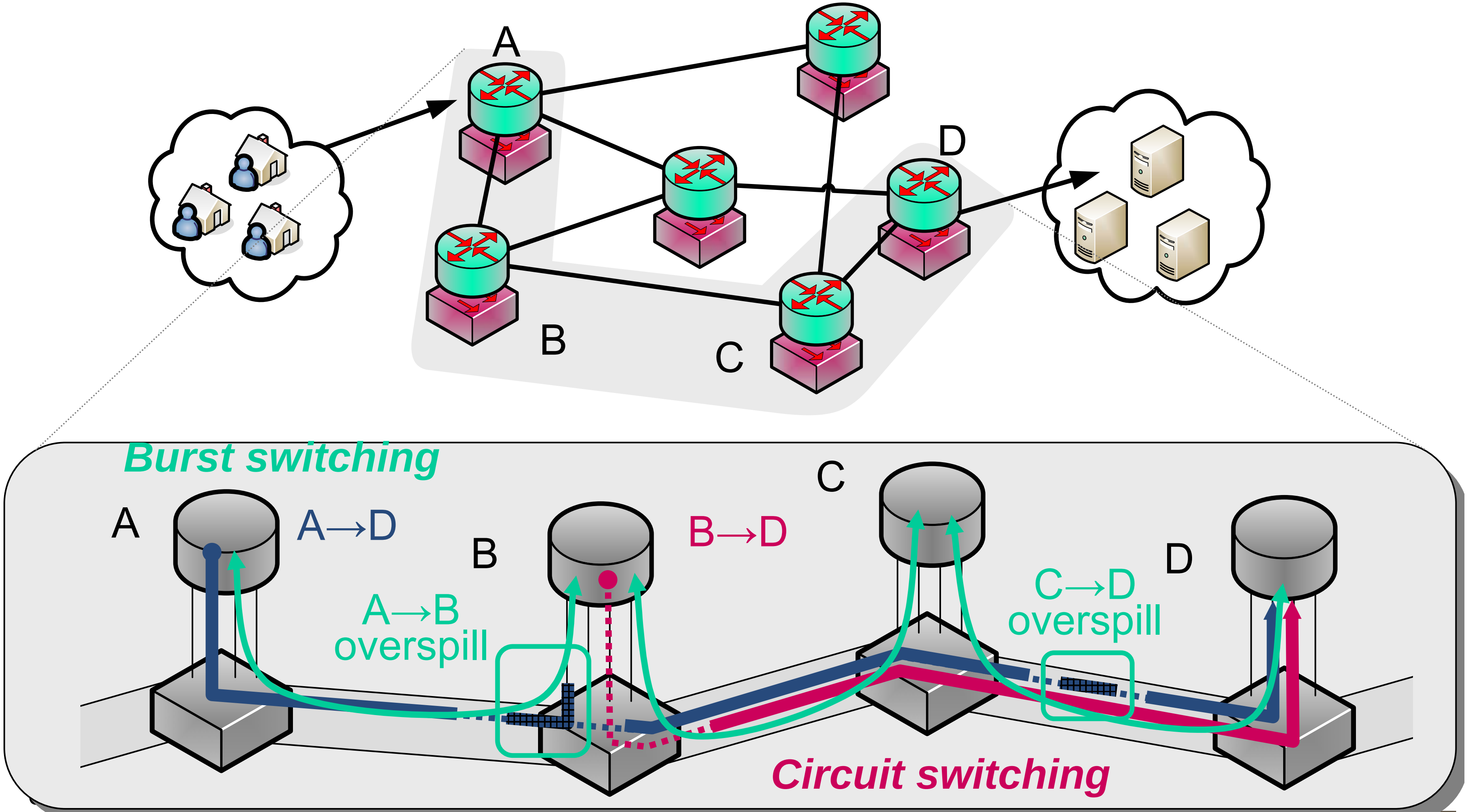
- **Hybrid architectures can offer a compromise**

- **Parallel: choice to either set-up OCS circuit between source & destination, or use OBS**
  - Note: can be overlay, where OBS makes use of OCS connections between OBS nodes



Note: CHEETAH [15] proposes a similar approach with IP and SONET as parallel layers

- **Overspill Routing In Optical Networks [8]:**



Differences with "classical" OBS:

- Anycast routing: user generally doesn't care where job is executed
- Burst starvation: not only network contention, also Grid resource contention
- Future reservation†: some jobs have very loose response time requirements, others are known long beforehand

†: note that current control planes such as GMPLS do not allow this (yet)

Introduction

Network Architecture

Routing

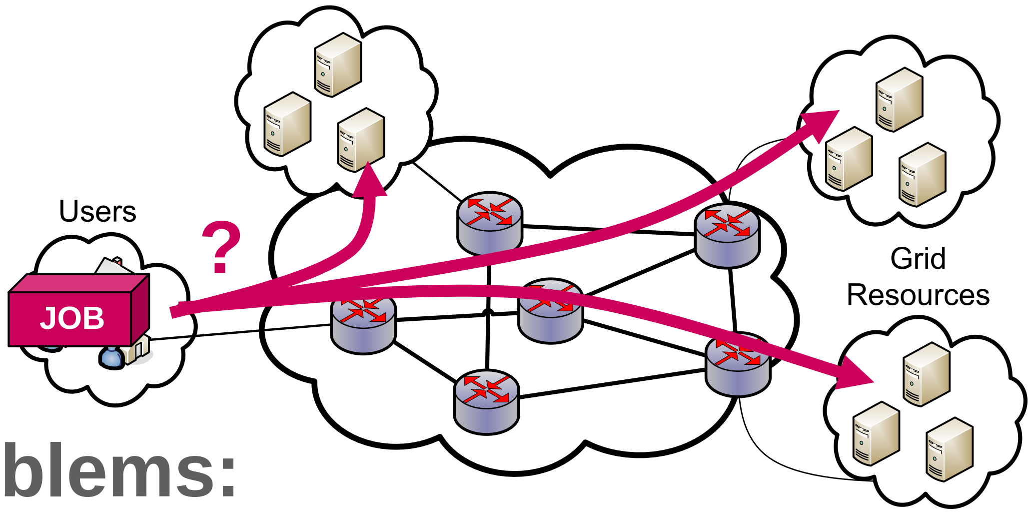
Dimensioning

Control Plane

Conclusions

Problem:

- Given a job, submitted by a user to an anycast address
- Find a set *r* containing at least one (and preferably one) suitable Grid site location accepting such jobs



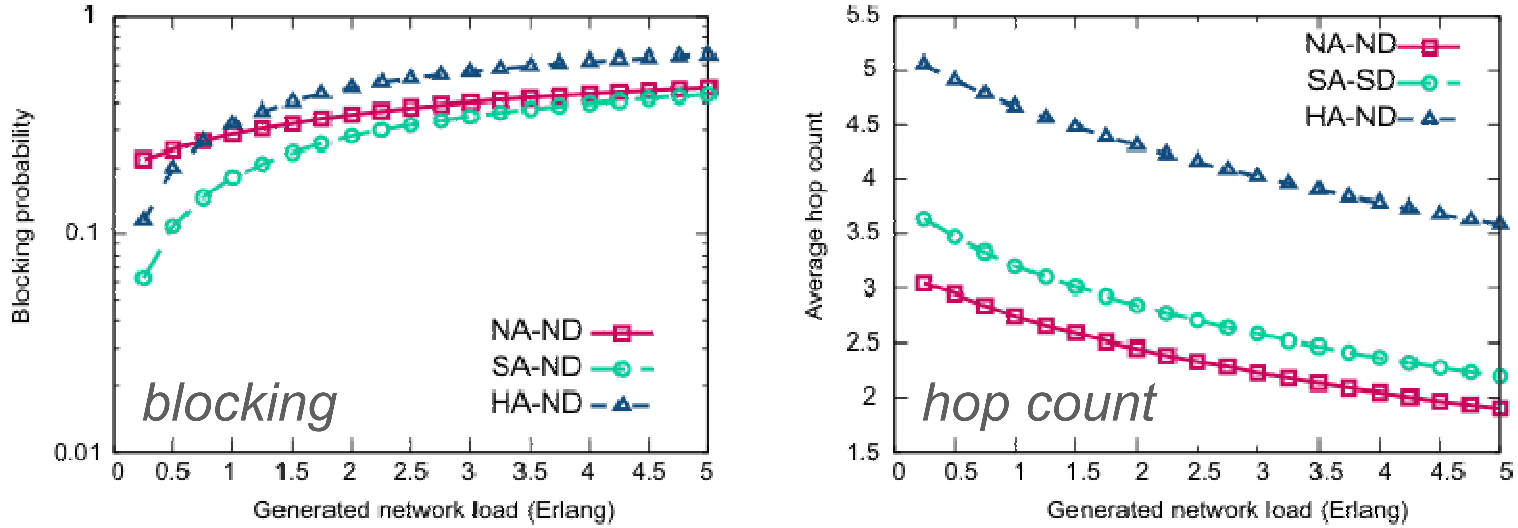
Sub-problems:

- Routing/deflection strategies
- Distributed multi-constrained routing algorithms

Routing Strategies

- **Soft Assignment (SA):**
  - Select a single destination node D (random, or some weighted function)
  - Other nodes along the path to D may execute job; or alter the destination to D’ to solve contention or starvation (→ deflection)
- **Hard Assignment (HA):**
  - Same selection as SA, but no modification (→ unicast)
- **No Assignment (NA):**
  - No explicit destination is chosen, but burst is passed on until a free Grid resource is found, or a pre-set slack time has expired

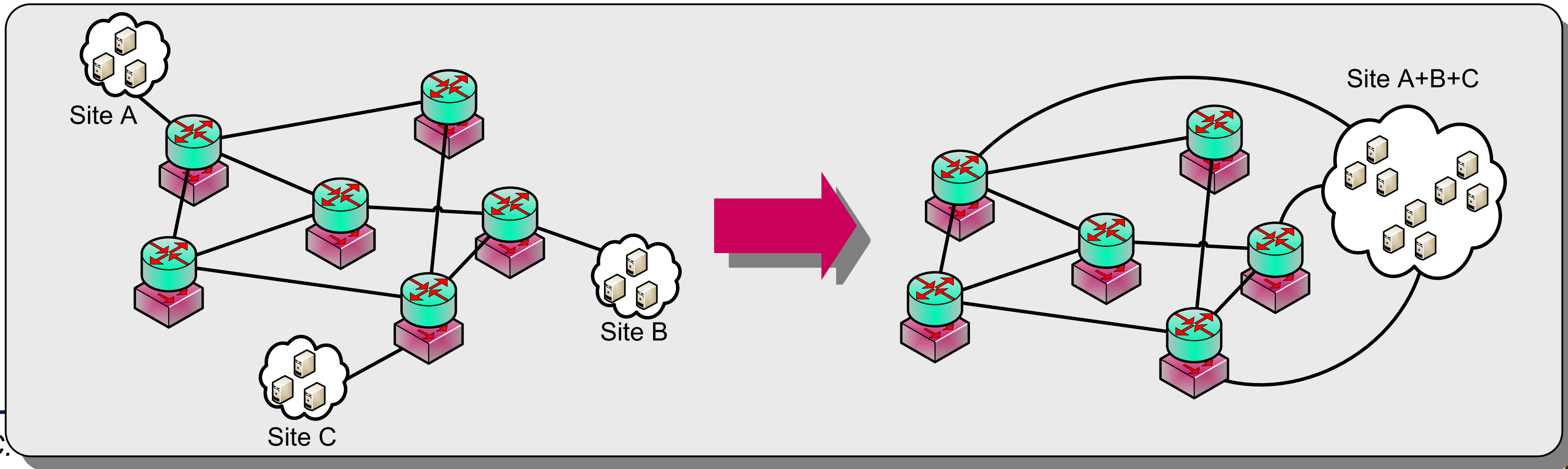
- **Soft Assignment** **performs best (least blocking)**
- **No Assignment** **outperforms** **HA** **for bigger loads**



For details and more results, see [10], available from
http://www.ibcn.intec.ugent.be/css_design/research/publications/index.php

Anycast SAMCRA

- **Problem:**
  - Incorporation of other metrics than just Grid resource availability leads to a multiple-constraint anycast routing problem
    (unicast multiple-constraint is already NP-complete)
- **Our solution:**
  - Introduce virtual topology to translate to unicast

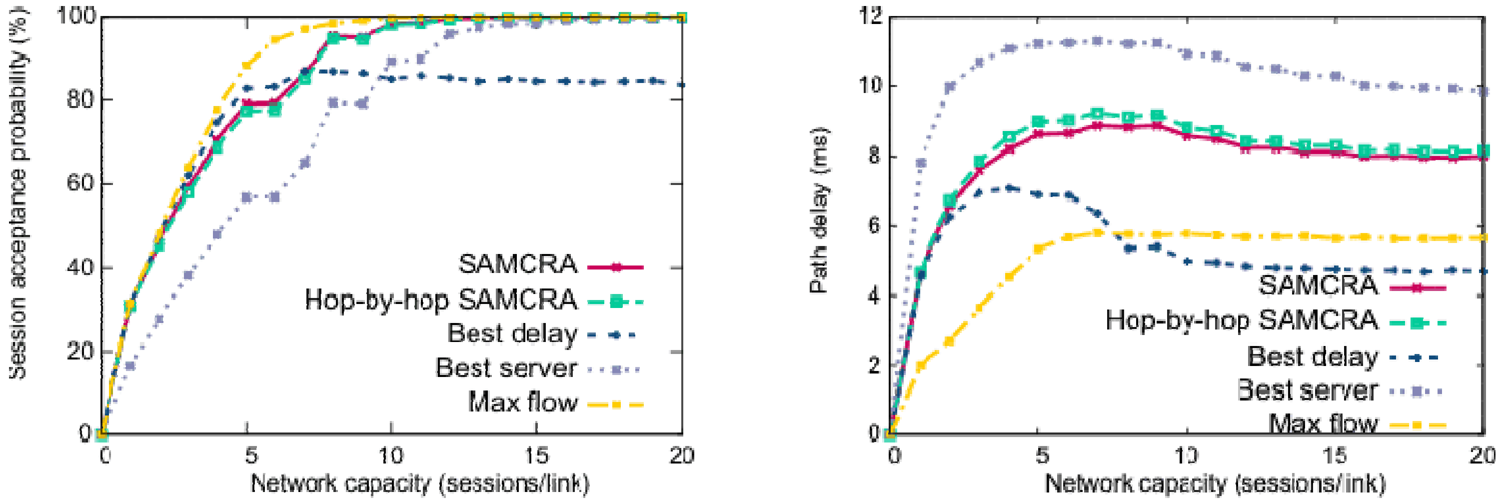


Anycast SAMCRA

- **Problem:**
  - Incorporation of other metrics than just Grid resource availability leads to a multiple-constraint anycast routing problem
    (unicast multiple-constraint is already NP-complete)
- **Our solution:**
  - Introduce virtual topology to translate to unicast
  - Use a Self-Adaptive Multiple Constraint Routing Algorithm (SAMCRA)
  - Use a novel path ordering avoiding sub-optimality and loops [11]

Anycast SAMCRA: results

- **Comparison with a Maximal-Flow upper bound shows that even distributed SAMCRA comes very close to (pseudo-)optimal acceptance rate**
- **Simpler heuristics, taking only 1 measure into account, do not come as close**



For details and more results, see [11]

UNIVERSITEIT GENT

FACULTY OF ENGINEERING

Introduction

Network Architecture

Routing

Dimensioning

Control Plane

Conclusions

IBBT

Case Study: excess load [9]

- **Network dimensioning for excess load**

- **Assuming**
  - Jobs arrive according to a Poisson process
  - Each Grid site is dimensioned for a steady-state load
  - A single site at a time suffers from excess load
  - This excess is offloaded to *k* other Grid sites

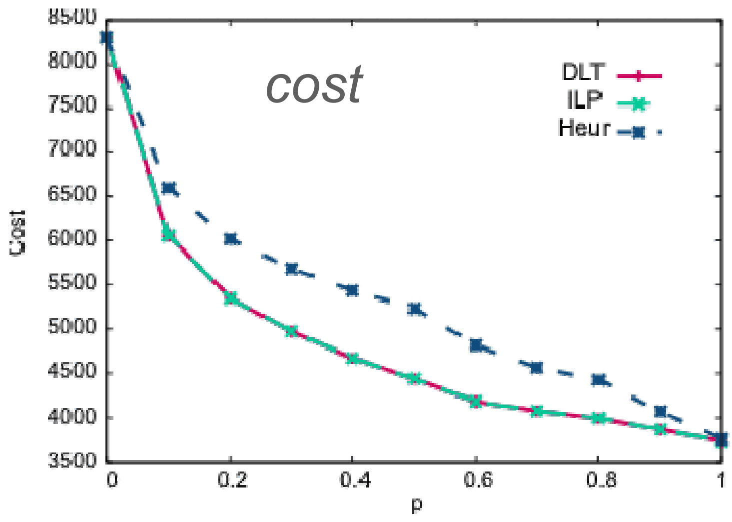
- **Find**
  - The minimal <u>network</u> dimension that can cater for each of the individual grid site overload scenarios

Methodology

- **For each scenario: generate series of jobs**
- **Integer Linear Programming (ILP):**
  - Per-job decision variable on which site to execute it
  - Global ILP solution over all overload scenarios
- **Heuristic:**
  - As ILP, but only solve individual scenarios (in parallel)
  - Take max. network dimensions over all scenarios
- **Divisible Load Theory (DLT):**
  - Real-value relaxation: workload is assumed to be arbitrarily divisible (total load = aggregate of all jobs)

Results

- **Cost vs. average connectivity for random 13-node networks:**



For details and more results, see [9], available from http://www.ibcn.intec.ugent.be/css_design/research/publications/index.php

- **Conclusion:**
  - **DLT** very close to optimal **ILP** solution, far more scalable
  - Heuristic scales even better, but results of less quality

Conclusions

- **Architecture:**
  - OBS seems a very promising candidate
  - Especially if it can be integrated with OCS in a hybrid form
- **Routing**
  - Anycast routing requires deployment of new algorithms
- **Excess load dimensioning algorithm**
- **Still many research opportunities**

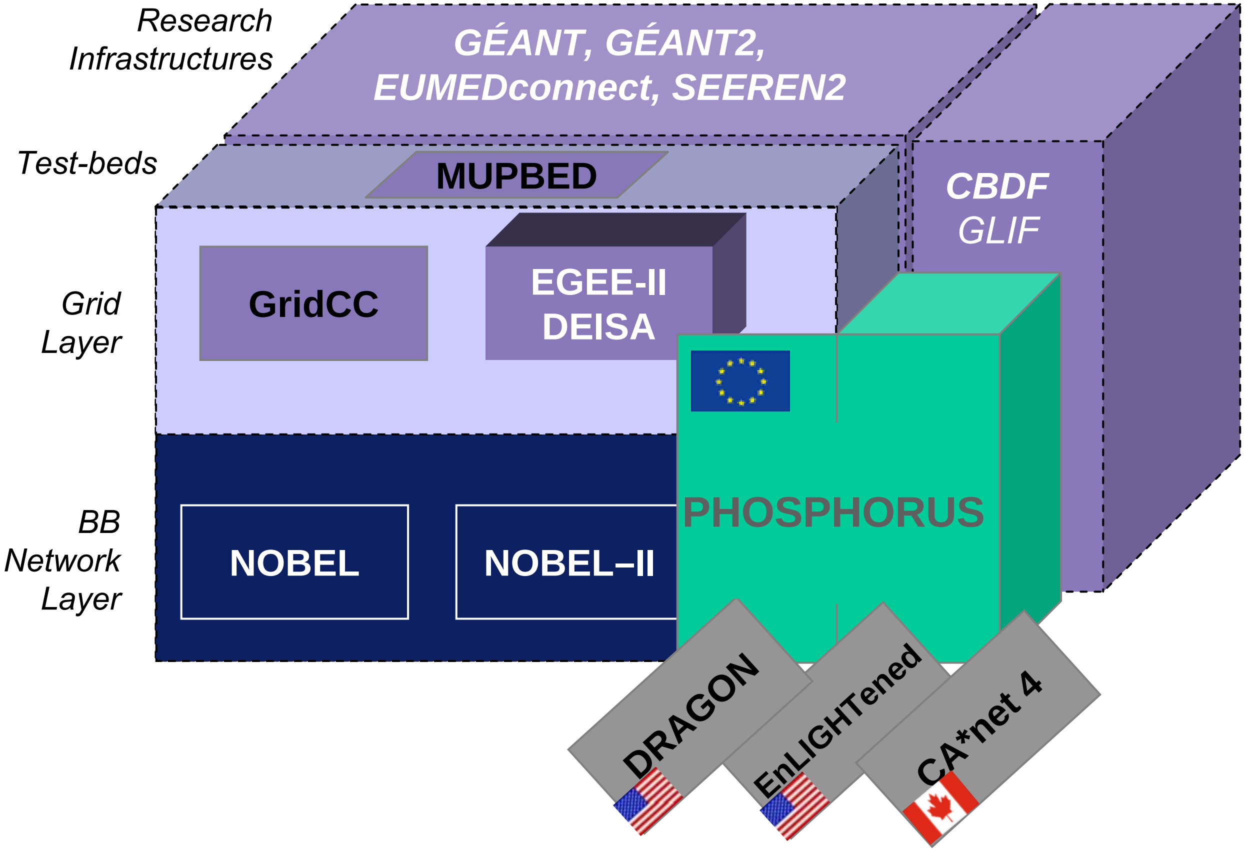
Challenges

- **Integrated OCS/OBS/hybrid control plane**
  - Interworking, migration…
- **Anycast OBS vs OCS?**
  - Performance comparison: job acceptance rate, response times, network utilization, overhead,…
- **Resilience**
  - Job migration, protection/restoration approaches…
- **Standardisation**
  - E.g. GoOBS architecture, burst format, routing protocols, inter-domain routing

Future work

- **Dimensioning algorithms**
- **Hybrid OBS/OCS architectures**
- **Resilience [19]:**
  - Fault management
  - Protection and restoration
- **Control plane**
- **Security and authentication**

- **Phosporus = new European optical Grid project, official start date 1 Oct. 2006 (aka 'Lucifer' [20])**



- **Phosporus will interact with:**
  - GÉANT2 (GN2 JRA3, JRA1 & JRA 5)
  - International activities: DRAGON, EnLIGHTened
- **Possible relationships with other EU projects**
  - focused on network layer technologies: NOBEL 1 & 2, EuQoS
  - focused on Grid layer: EGEE-II, GridCC
  - test-bed oriented: MUPBED


OPTICAL
GRID
AHEAD
That's all folks!
… any questions?

References 1-8

- [1] G. Fox, A.J.G. Hey, F. Berman, *Grid computing: making the global infrastructure a reality*, John Wiley & Sons, Mar. 2003, ISBN: 0-470-85319-0.
- [2] *LCG - LHC Computing Grid Project*, http://lcg.web.cern.ch/LCG/
- [3] M. De Leenheer, et al., *A View on Enabling Consumer Oriented Grids through Optical Burst Switching*, IEEE Commun. Mag., Mar. 2006, pp. 124-131.
- [4] D. Simeonidou, et al., *Dynamic optical-network architectures and technologies for existing and emerging grid services*, J. of Lightwave Techn., Vol. 23, No. 10, Oct. 2005, pp. 3347–3357.
- [5] J. Baert, et al., *Hybrid optical switching for data-intensive media grid applications*, Proc. Workshop on Design of Next Generation Optical Networks, Ghent, Belgium, 6 Feb. 2006, pp. 9-14.
- [6] C. Develder, et al., *Node architectures foroptical packet and burst switching*, Tech. Digest Int. Topical Meeting on Photonics in Switching (PS2002), (invited) paper PS.WeA1, Cheju Island, Korea, 21-25 Jul. 2002, pp. 104-106.
- [7] M. Veeraghavan, et al., *On the Use of Connection-Oriented Networks to Support Grid Computing*, IEEE Commun. Mag., Mar. 2006, pp. 118-123.
- [8] E. Van Breusegem, et al., *Overspill routing in optical networks: A true hybrid optical network design*, IEEE J. Selected Areas in Commun., Apr. 2006, pp. 13-26.

References 9-14

- [9] P. Thysebaert, et al., *Using divisible load theory to dimension optical transport networks for grid excess load handling*, Proc. Int. Conf. on Autonomic and Autonomous Systems & Int. Conf. on Netw. (ICAS/ICNS 2005), Papeete, Tahiti, 23-28 Oct. 2005.
- [10] F. Farahmand, et al., *A multi-layered approach to optical burst-switched based grids*, Proc. of Workshop on Optical Burst/packet Switching (WOBS2005), held on Broadnets 2005, 2nd Int. Conf. on Broadband Commun., Netw. and Sys.net, Boston, USA, 3-7 Oct. 2005, pp. 127-134.
- [11] T. Stevens, et al., *Distributed Job Scheduling based on Multiple Constraints Anycast Routing*, accepted for Broadnets 2006.
- [12] P. Szegedi, et al., *Signaling Architectures and Recovery Time Scaling for Grid Applications in IST Project MUPBED*, IEEE Commun. Mag., Mar. 2006, pp. 74-82.
- [13] T. Lehman, et al., *DRAGON: A Framework for Service Provisioning in Heterogeneous Grid Networks*, IEEE Commun. Mag., Mar. 2006, pp. 84-90.
- [14] I.W. Habib, et al., *Deployment of the GMPLS Control Plane for Grid Applications in Experimental High-Performance Networks*, IEEE Commun. Mag., Mar. 2006, pp. 65-73.

References 15-20

- [15] X. Zheng, et al., *CHEETAH: Circuit-switched high-speed end-to-end transport architecture testbed*, IEEE Commun. Mag., Aug. 2005
- [16] I. Foster, et al., *The Physiology of the Grid An Open Grid Services Architecture for Distributed Systems Integration*, Globus Draft, Jun. 2002, available from http://www.globus.org/ogsa/
- [17] J. Recio, et al., *Evolution of the User Controled Lightpath Provisioning System*, Proc. 7th Int. Conf. on Transparent Optical Networks (ICTON), Barcelona, Jul. 2005.
- [18] –, *Application Brief: Dynamic Resource Allocation Controller (DRAC)*, available from www.nortel.com/solutions/optical/collateral/nn-110181-1130-04.pdf
- [19] J.P. Vasseur, M. Pickavet, P. Demeester, *Network Recovery / Protection and Restoration of Optical, SONET-SDH, IP, and MPLS*, Morgan Kaufman, Aug. 2004, ISBN: 0-12-715051-X.
- [20] N. Ciulli, *Grid services enabled photonic infrastructure in Europe*, Int. Workshop on the Future of Optical Networking, held at OFC 2006, Anaheim, CA, USA, Mar. 2006

Note: see http://www.ibcn.intec.ugent.be/css_design/research/publications/ for our own publications