Research Challenges for Optical Grid Computing

M. De Leenheer, C. Develder, T. Stevens, J. Vermeir, F. De Turck, B. Dhoedt, P. Demeester

OGF20, Manchester, UK
May 9, 2007
Introduction (1)

- **eScience:**

  - By 2015 it is estimated that **particle physicists** will require exabytes \((10^{18})\) of storage and **petaflops** per second of computation.

  - CERN’s LHC Computing Grid (LGC) will start operating in 2007 and will generate **15 petabytes** annually (that’s ~2Gbit/s).

  - LHC = Large Hadron Collider: particle accelerator
  - 50 CDROMs = 35 GB

  - CD stack with 1 year LHC data (~ 20 km or 12.5 mi)
  - Balloon (30 km or 18.6 mi)
  - Concorde (15 km or ~9.3 mi)
  - Mt. Blanc (4.8 km, or 3 mi)
Consumer service:

- Eg. video editing: 2Mpx/frame for HDTV, suppose effect requires 10 flops/px/frame, then evaluating 10 options for 10s clip is **50 Gflops** (today’s high performance PC: ~10 Gflops/s)

Online gaming:  
e.g. Final Fantasy XI: 1,500,000 gamers

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

Multimedia editing
- **Grid** opportunities ranging from academia over corporate business to home users
- **Optical** data speeds \( \geq \) internal PC bus speeds ⇒ network speed no bottleneck
Introduction

Network Architecture

Routing

Multiple Domains

Conclusions
**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**
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

†: e.g. 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 (hardware)
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
Overspill Routing In Optical Networks:

- Hybrid OBS/OCS: ORION

**Burst switching**
- A → D
- A → B overspill

**Circuit switching**
- B → D
- C → D overspill
Introduction
Network Architecture
Routing
Multiple Domains
Conclusions
Optical Grid specifics

- Differences with “classical” optical networks or “classical” Grids:
  - **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)
Problem Statement

- **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
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
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
Anycast SAMCRA: results

- Comparison with a **Maximum-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

---

*T. Stevens et al., “Distributed Job Scheduling based on Multiple Constraints Anycast Routing, Broadnets 2006*
Introduction
Network Architecture
Routing
Multiple Domains
Conclusions
Grid applications pose challenging demands:

- Connections extend to application end-points (rather than traditional network elements)
- **On-demand** bandwidth provisioning, both immediate and advance reservations
- Very **dynamic** use of end-to-end networking resources
- Requires **(near) real-time** feedback for signaling and provisioning
- **Heterogeneous** network, computing and storage resources in **multiple domains**
- **Diversity** in holding times and bandwidth granularity
- …
Multiple Domains: approaches

- Centralized
  - Bad scaling behaviour
  - Global view allows optimal decisions

- Fully distributed
  - High number of control plane events
  - Authority remains at end nodes

- Proxy approach
  - Improve scaling by limiting domain scope
  - Aggregate control plane events
Anycast proxy infrastructure

Clients and resources only use anycast communications

Control plane traffic reduction through state aggregation

Authority is shared by client and server proxies
T. Stevens et al., “Distributed Service Provisioning Using Stateful Anycast Communications, IEEE LCN 2007 (Submitted)
Introduction
Network Architecture
Routing
Multiple Domains
Conclusions
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

- Multiple Domains

- Still many research opportunities...
Challenges

- Integrated OCS/OBS/hybrid control plane
  - Interworking, migration, node architecture, …

- Dimensioning and network planning

- Resilience
  - Job migration, protection/restoration approaches…

- Standardisation
  - E.g. GoOBS architecture, burst format, routing protocols, inter-domain routing
Questions?
Phosphorus = new European optical Grid project, official start date 1 Oct. 2006 (aka ‘Lucifer’)

Phosphorus 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