MetaCentrum

Virtualized Grid on a Virtualized Network – Magrathea, VirtCloud, and SBF

Luděk Matyska

CESNET & Masaryk University

Prague & Brno

Czech Republic

Grid Virtualization

- **Ability to set up *virtual clusters* spanning physical resources of a wide area grid**
  - Virtual clusters are fully independent, up to the Layer 2 of the underlying network
  - Temporary and permanent clusters
- **Each such cluster runs user specified images including any Grid middleware user need**
  - Grid operations support directly only limited number of images and grid middleware systems
- **Two levels of scheduling**
  - To schedule the virtual cluster (nodes and network)
  - To schedule jobs within virtual cluster

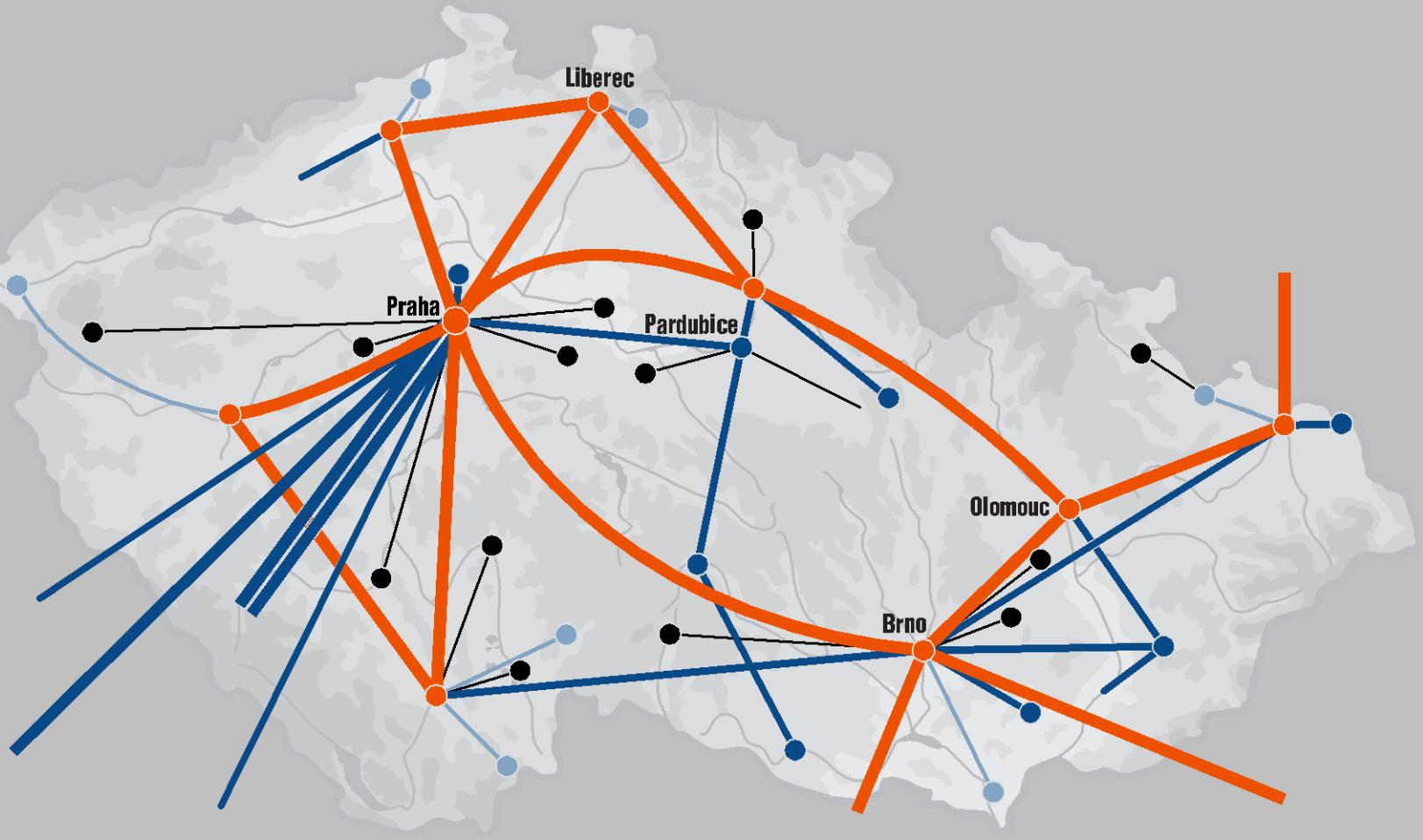
MetaCentrum Environment

- **Core of the Czech National Grid infrastructure**
- **Some numbers**
  - 8 sites (CESNET plus universities)
  - >1300 cores + >1500 cores for EGEE/WLCG
  - ~0,2 PB on disks and 0,4 PB on tapes
- **Based on clusters coordinated via PBSPro**
  - Global and individualized user/team queues
  - Global storage provided by AFS and NFSv4
  - Site storage via NFS
  - /scratch systems on individual hosts

CESNET network – CESNET2

- **Backbone based on multi 10 Gbps lines**
  - Some of them on dark fiber with direct access to optical level
- **Complemented with 1 Gbps lines to smaller sites**
- **Runs MPLS on a country-wide scale**
- **Dedicated 10 Gbps lines (physical and lambdas) available between Brno, Prague, and Pilsen**
  - Prague—Pilsen below 150 km
  - Prague—Brno below 250 km
- **10 Gbps to GEANT2**
- **Dedicated experimental 10 Gbps independent lines to Amsterdam and Chicago**

CESNET2 Network Topology



MetaCentrum Virtualization

- **More than half of nodes runs Xen system**
  - 16 core nodes still using VServer
- **User jobs run in virtual machines (DomU)**
  - Access to Dom0 restricted to administrators
  - Very limited impact on compute intensive jobs
    - Usually below 3% compared to physical machine
  - Moderate impact on data intensive jobs
    - Needs careful scheduling of data intensive jobs
    - Using 1-2 processors for data processing on 8 core nodes

Images

- **Two major classes directly supported**
  - The native MetaCentrum environment
    - OpenSuSe and Debian
  - The EGEE/gLite environment
  - Each node “runs” both environments in separate DomU virtual machines
- **Experimental (pre-production) support for user supplied images**
  - Needs careful security considerations

Scheduling

- **PBSPro does not directly support virtualized environment**
- **Magrathea – a scheduler extension to deal with multiple virtual machines on physical hosts**
  - Takes care of which resources are visible
  - Deals with special states of virtual machines
    - E.g. Hibernated or Frozen
  - Activates/Deactivates the scheduled, suspended and finished images

Work with images

- **Each node runs two virtual machines**
  - Both in a minimized state, but active
  - Supports fast switch between environments
  - *Booot* to extend/shrink from/to minimized state
    - Also real boot for new images
- **Experimental support to upload new images**
  - Full boot, fresh environment
  - Hibernated and frozen images, allow to continue previously stopped work

Network Virtualization

- **Motivation:**
  - Security
    - Fully encapsulate potentially dangerous user's images
    - Use of weak security inside user's images
  - Decouple physical and user network view
    - Users can use their own IP ranges
    - Users can hardwire IP addresses into application
      - *Support parallel runs of such applications*
    - Traffic encapsulation

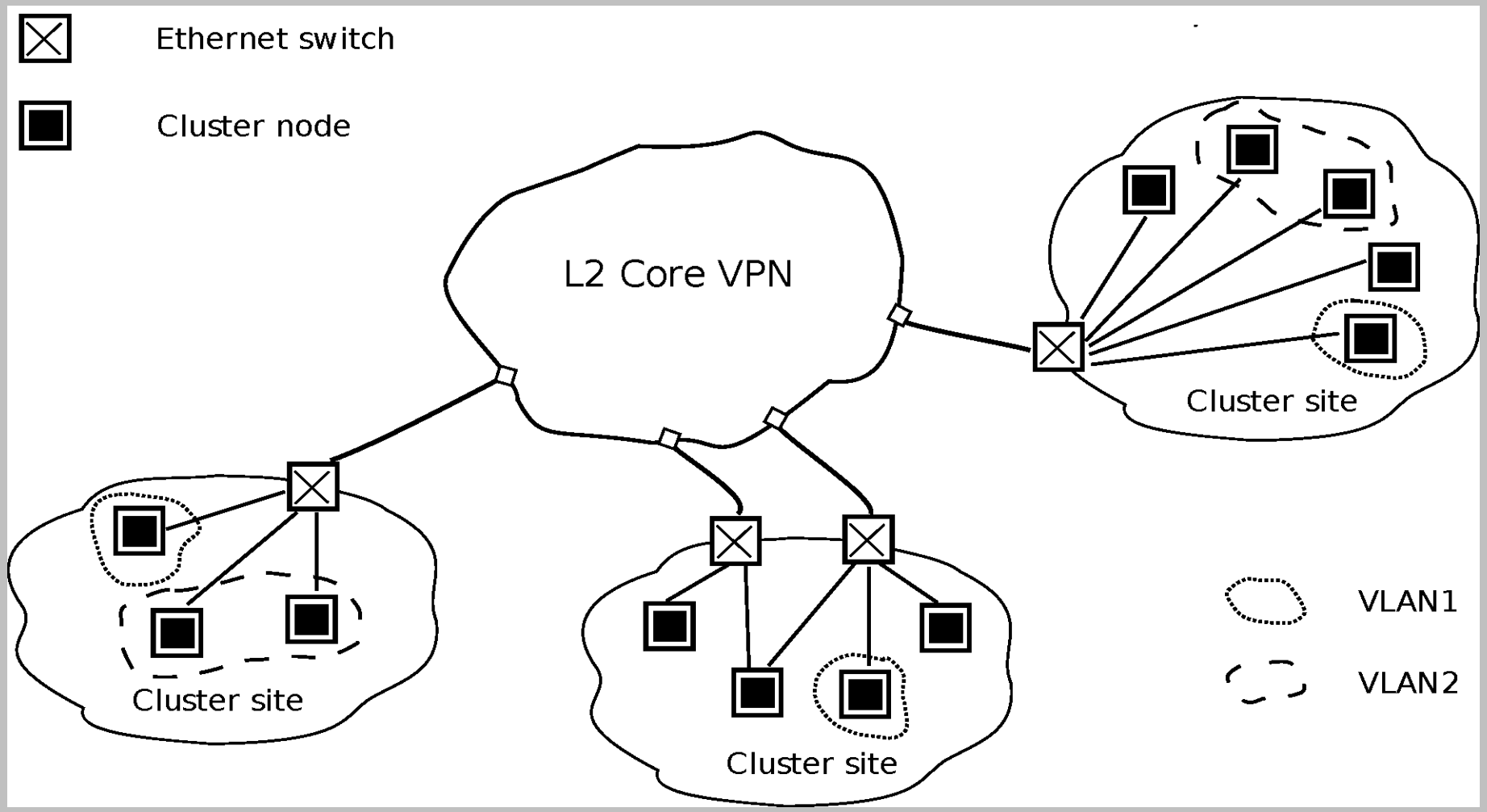
Network Virtualization – VirtCloud

- **Virtual Private Circuits/Networks (VPN)**
  - Provide encapsulation
  - Usually high overhead
  - User managed
- **Virtual LAN (VLAN) technology**
  - At the Layer 2 of the network
  - Encapsulation
  - Very low overhead
  - Administrator managed

Virtual VLAN

- **User managed Virtual LAN**
  - Needs support at the network physical level
  - Once set up by network administrators
  - Then manageable by users
  - QinQ, VLAN tagging, …
- **An intermediate provision for virtual grids**
  - Virtual VLAN managed by the infrastructure
  - Users see “their” network
  - No messing with other traffic (from the same or other users)

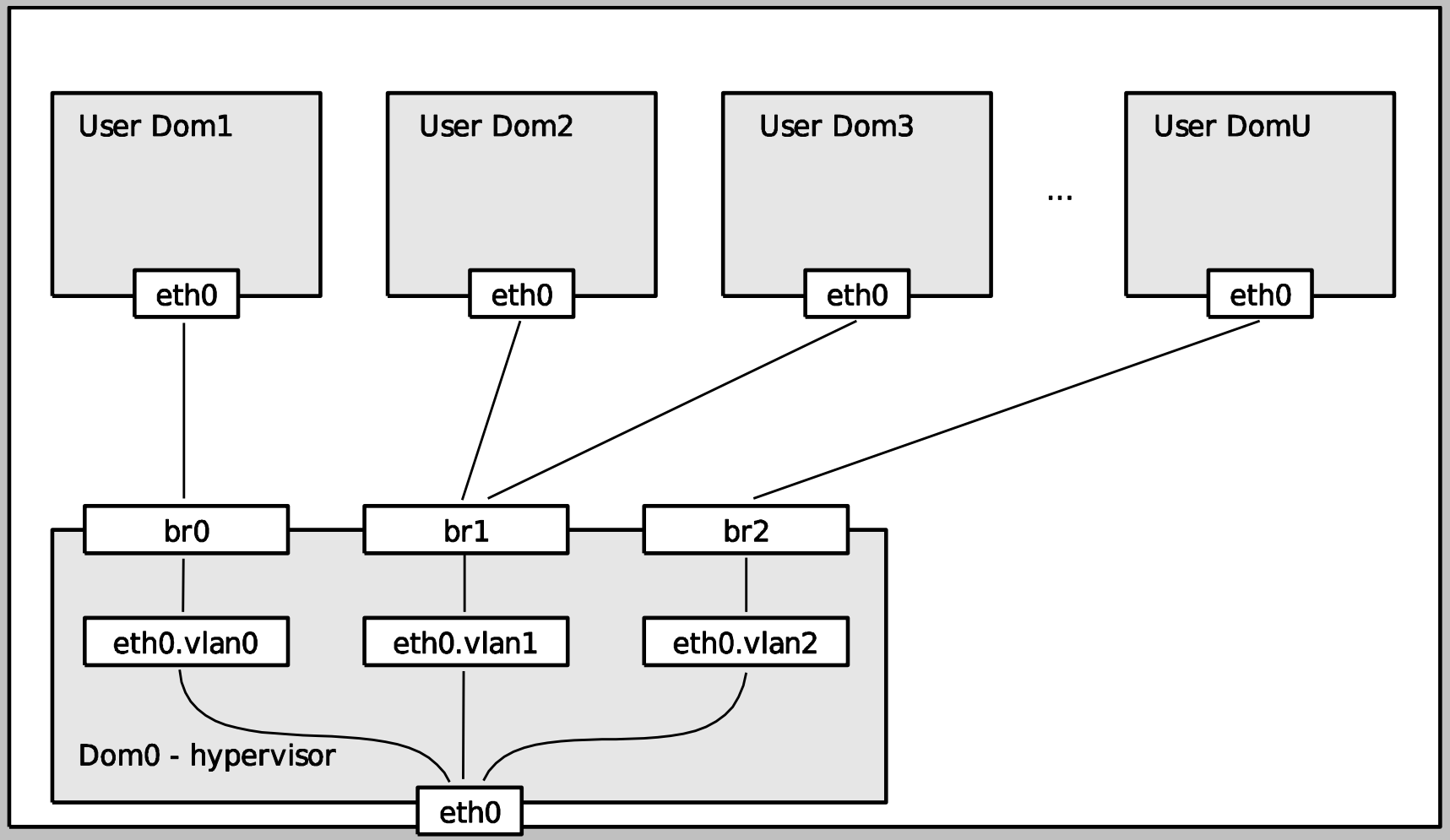
Network Connections



Host Configuration

- **All traffic goes through tagged VLANs**
  - Managed by specific scheduler
  - Set up in the Dom0 (privileged)
- **User traffic maps onto tagged VLAN**
  - Through virtual network interface in DomU
  - Users could have root privileges in their images
  - Mapping done by bridge out of user's control

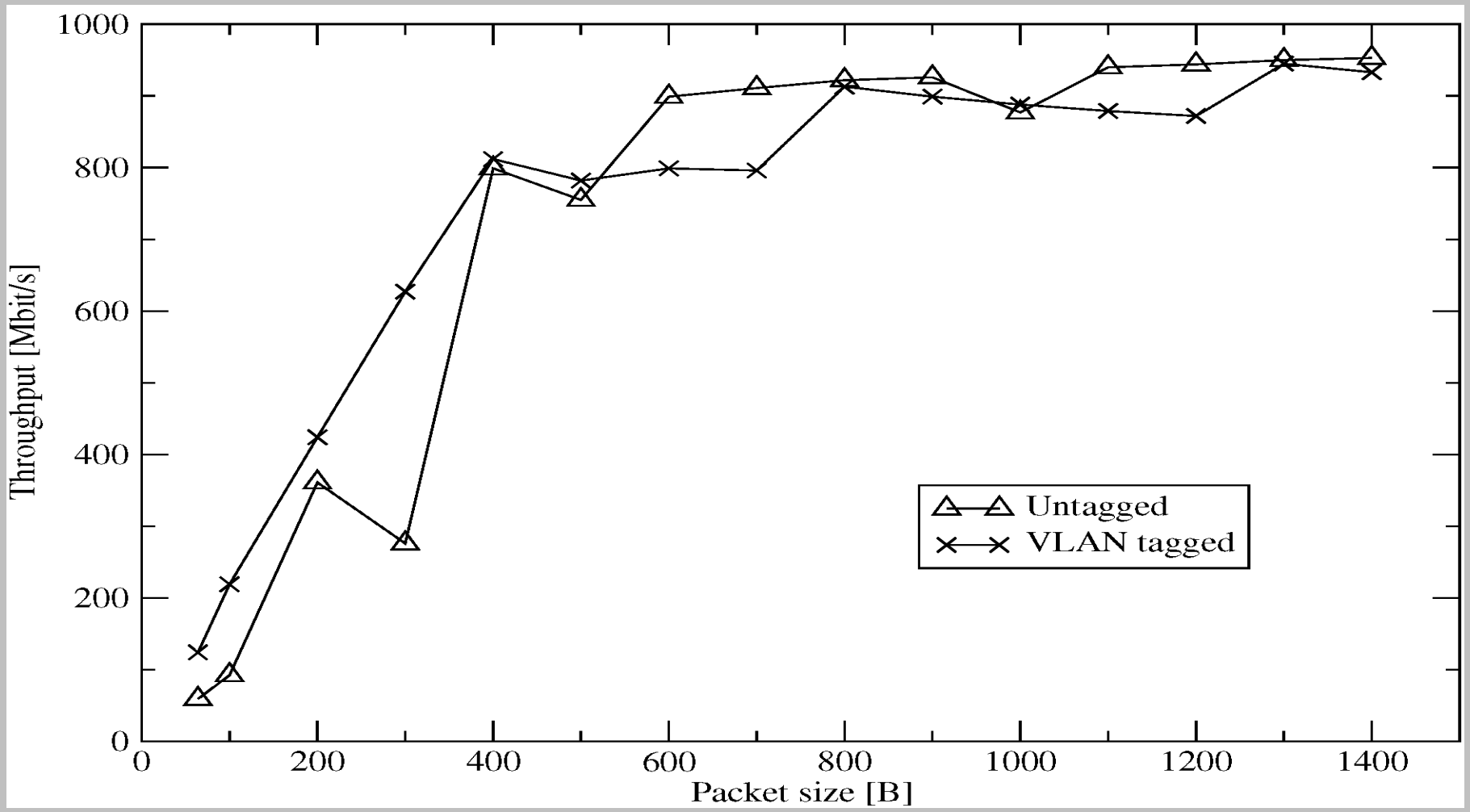
Host Configuration – Schema



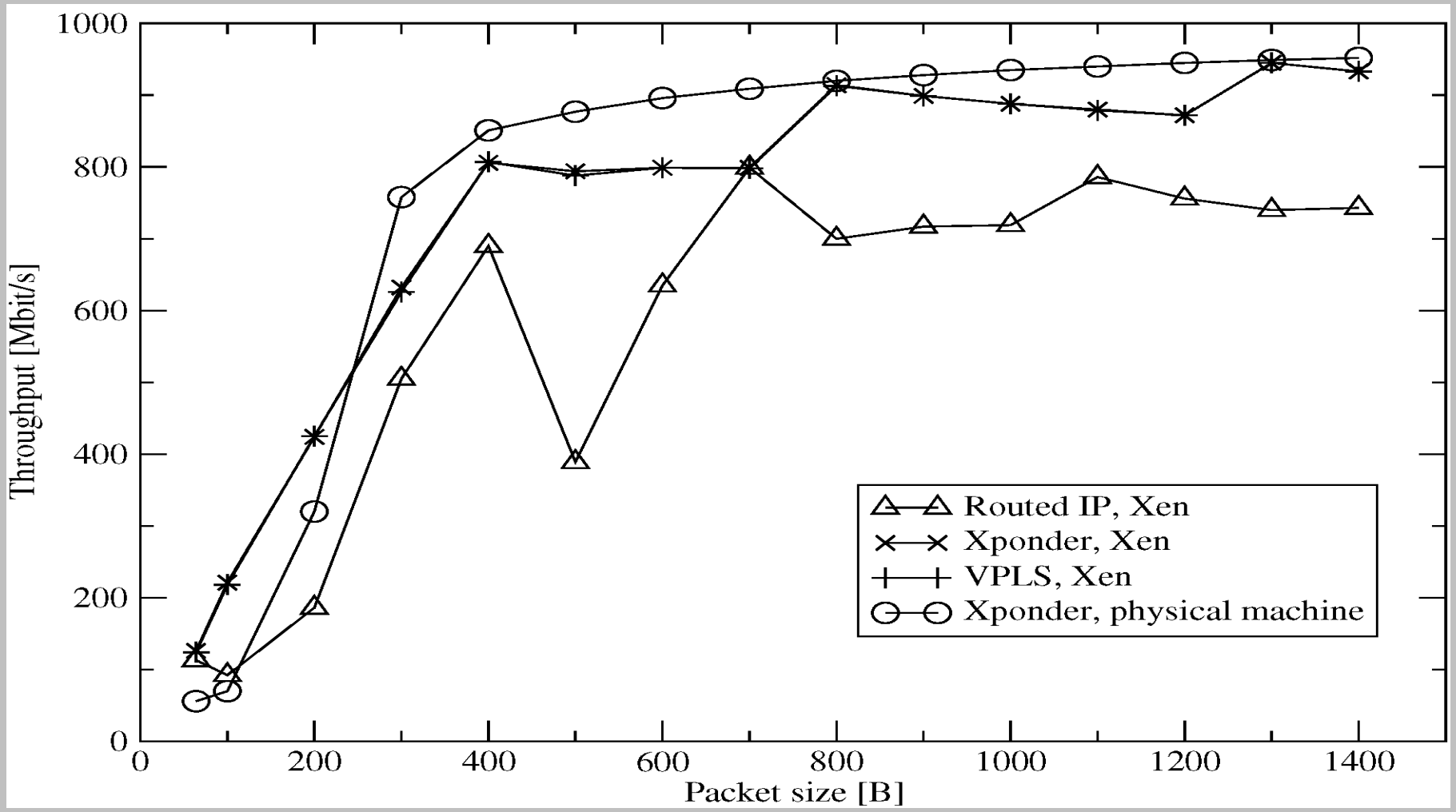
Network experiments

- **Need to evaluate impact of VLAN tagging and virtualization on network performance**
- **Experiments run a country wide network**
  - Between physical machines
  - Between virtual (DomU) machines
  - With and without tagging
- **Network modes**
  - MPLS/VPLS in the production network
    - 1 Gbps dedicated channel
  - XPonder based dedicated optical network
    - 10 Gbps over lambda
  - Routed IP without any dedication

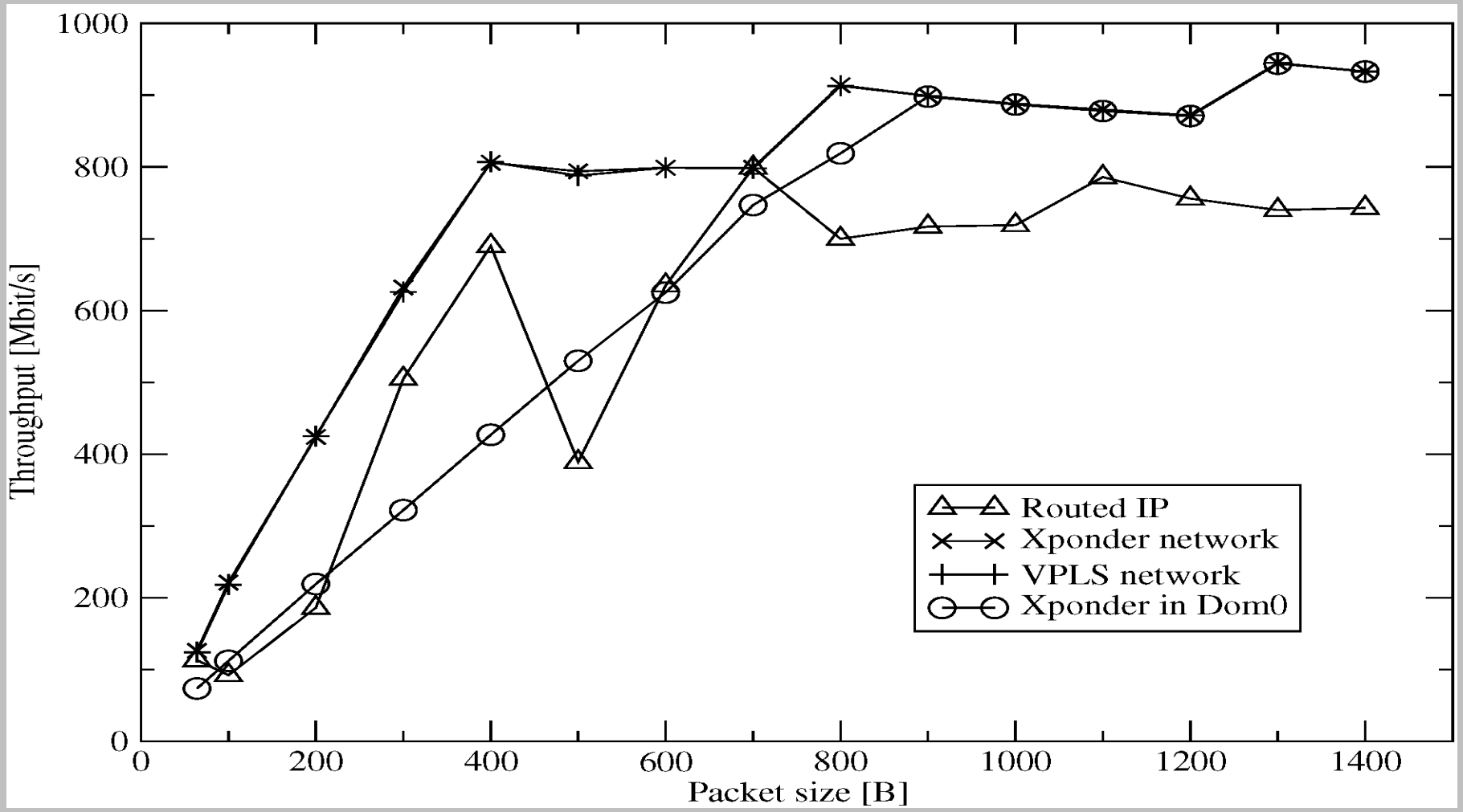
UDP, Price of VLAN tagging in Xen



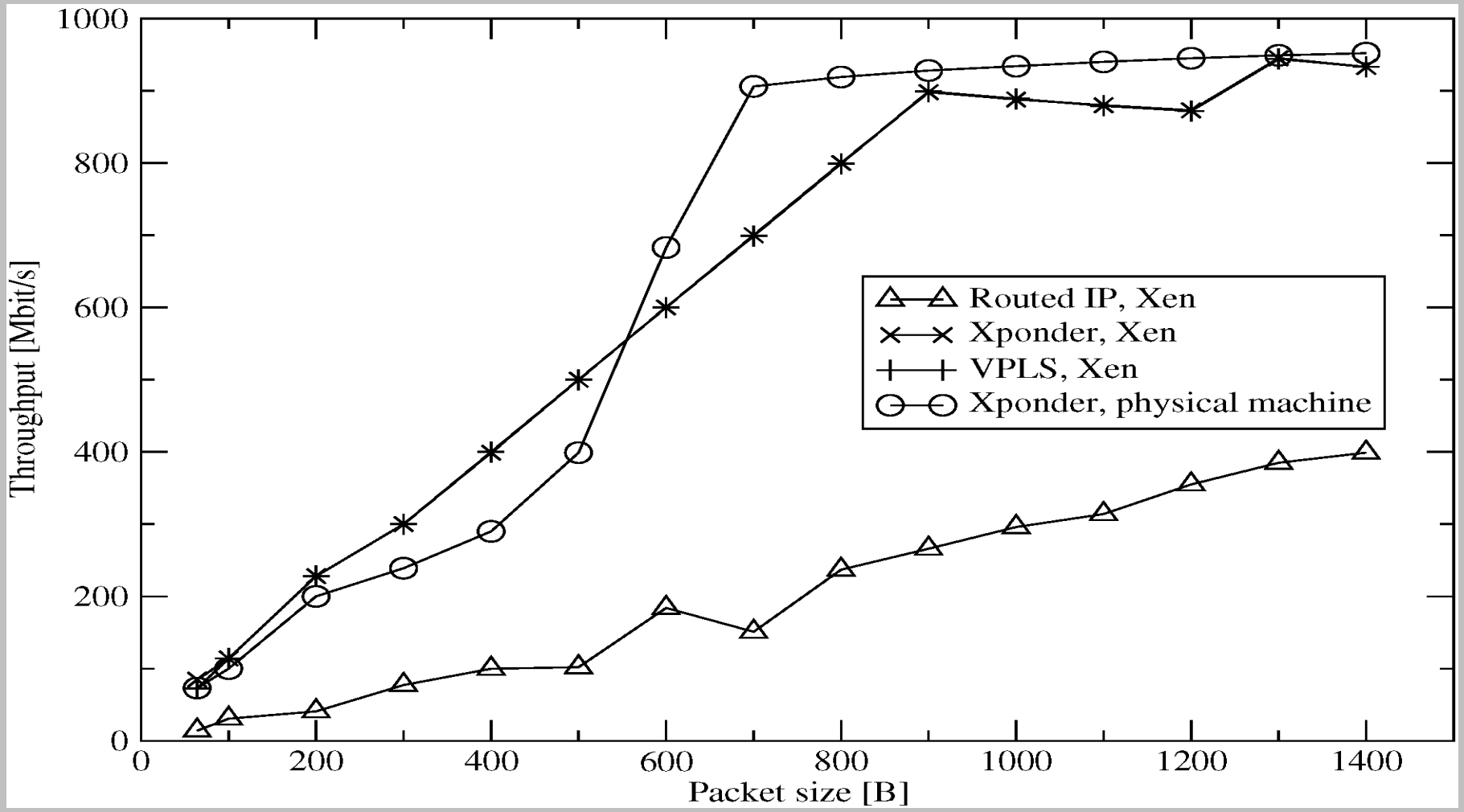
UDP, Brno – Prague, Dom0



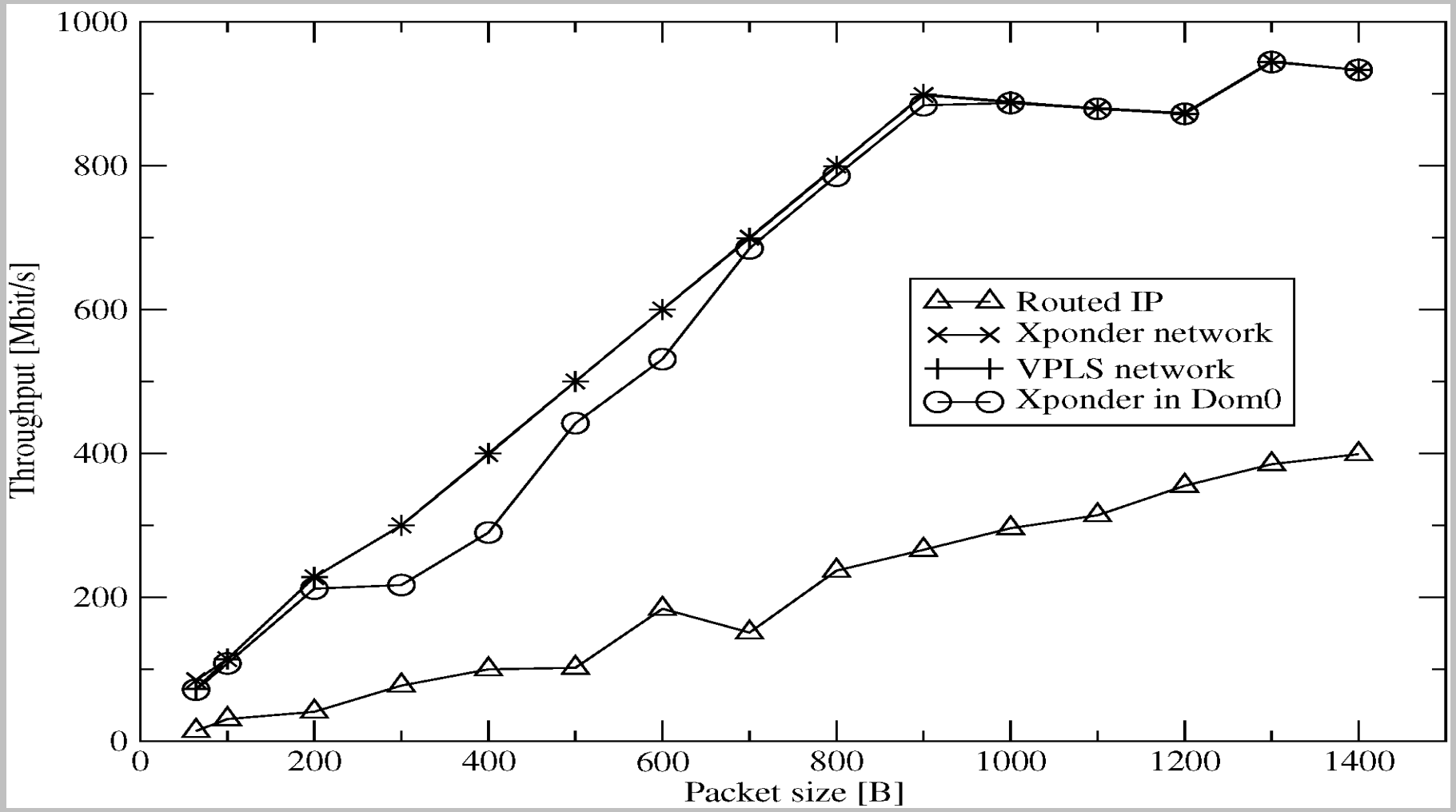
UDP, Brno – Prague, DomU



UDP Brno – Pilsen, Dom0



UDP Brno – Pilsen, DomU



SBF service

- **Management of the virtual VLANs**
- **Takes care of the virtual VLAN life cycle**
- **Called by PBSPro when a request for a virtual cluster is issued by user**
  - Could be implicit as part of e.g. an MPI job setup
- **Configures active network elements and returns VLAN tag/number**
- **PBSPro with Magrathea manages the host**
  - Configures bridging
  - Boots/activates the requested image

Access to Virtual Grids

- **Open network (possible) in case of certified images**
  - Routing provided by the infrastructure
- **Closed network otherwise**
  - Provide a gateway into the virtual grid
    - Authenticated, authorization part of the virtual cloud setup
    - Authorization under user's control
  - User responsible for all traffic to/from virtual grid

Summary

- **MetaCentrum provides full virtualized environment**
- **All important components in production or pre-production deployment on a country wide infrastructure**
  - Virtualized hosts
  - Virtual network
  - Scheduling: PBSPro, Magrathea, SBF
  - Network encapsulation: VirtCloud, VLAN tagging