Supercomputer Within Supercomputer: Discovery of Interconnect Topologies by Complete Supercomputing System Emulation

Łukasz P. Orłowski

A*STAR Computational Resource Centre
Dept. of Applied Mathematics and Statistics, Stony Brook University
Institute for Advanced Computational Science, Stony Brook University

Advisors:
Marek T. Michalewicz, Yuefan Deng

Why study interconnect topologies?

- “Moore’s law slows to a crawl” - Top500 newsletter 03/15/2017

- Choice of proper topology offers performance gain

Why virtual supercomputer?

- Flexibility of architecture adjustment
- Architectural invariants become variable parameters
- Software automation of interconnect rewiring
- Enabling interconnect modification while the computation is running
- Good tool to explore interconnect topologies

Requirements

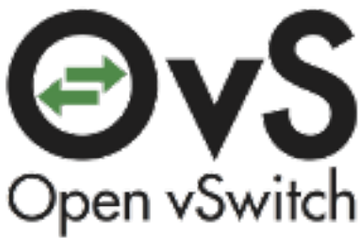
- Change of interconnect network topology is transparent for compute nodes
- Network topology may be changed when the system is running
- Network topology may be modified while computations are running on a system
- The system is operational even when the virtual interconnect network is down (of course then compute nodes cannot communicate)
- Compute nodes are stateless

What might be used to build a virtual supercomputer

- Virtual machines with virtual CPUs pinned to physical cores
- Virtual switch processes pinned to corresponding cores
- Virtual link latencies set to emulate cable lengths
- Network behaviour scenarios implemented with Software-Defined Networking (SDN)
- When deployed into large SMP, one may simulate… hundreds of compute nodes!

Software components

Open vSwitch (OVS)
Virtual switches



OpenFlow
SDN management protocol

Open Network Operating System (ONOS)
SDN controller

KVM + QEMU
virtualisation of nodes

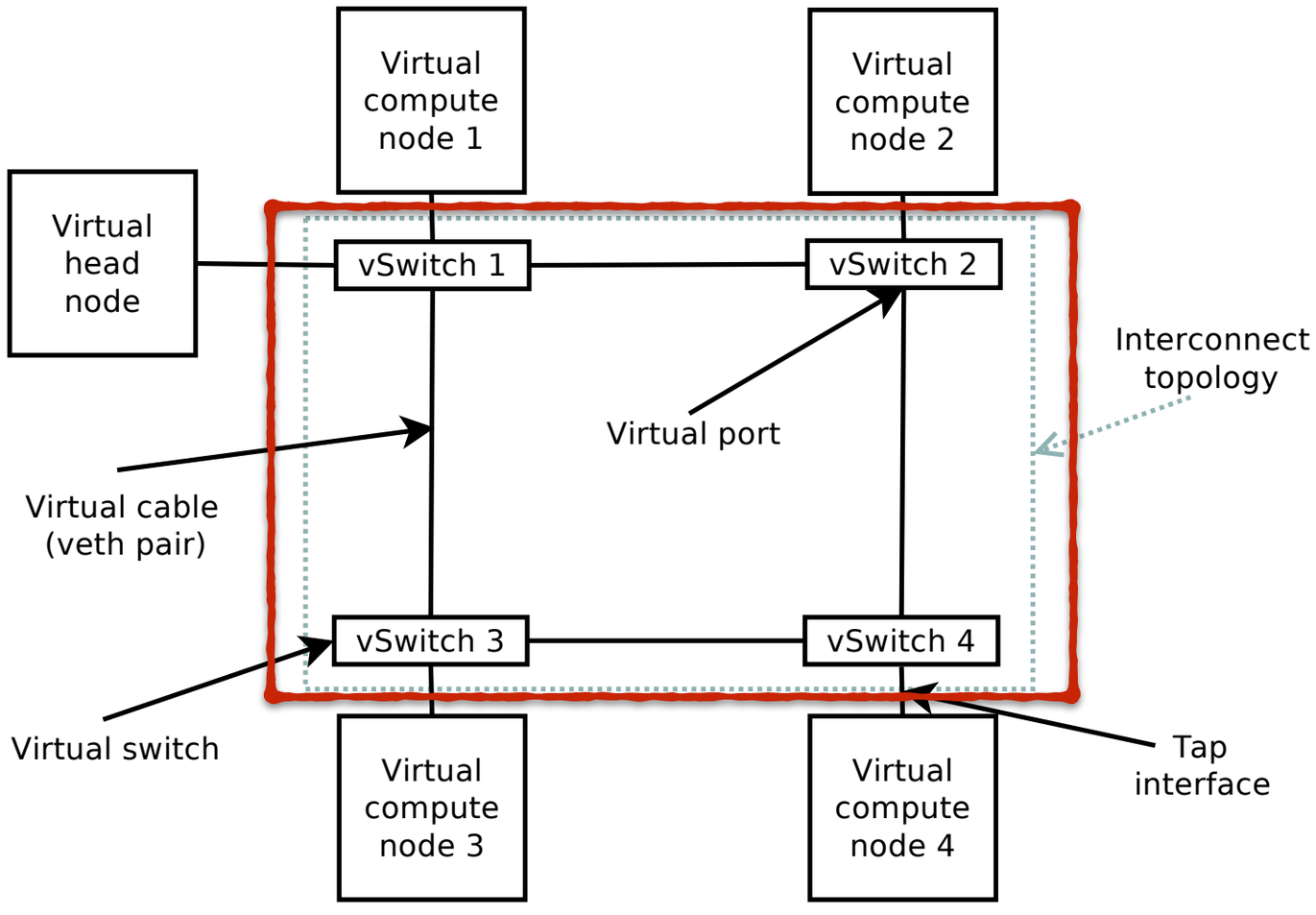
Libvirt
virtualisation toolkit and API

A*STAR Computational Resource Centre

Stony Brook University

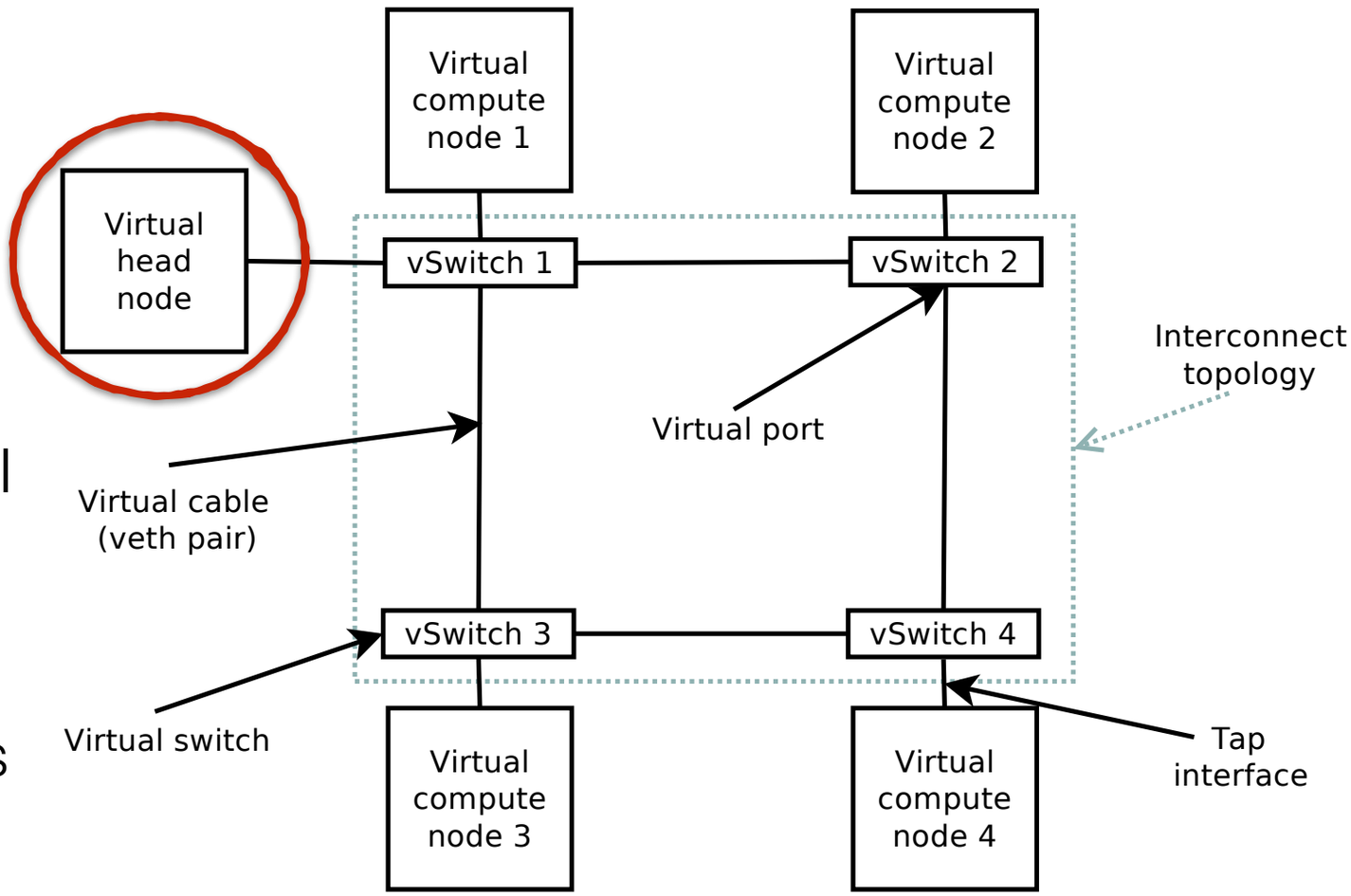
Virtual interconnect

- The most important component for topology emulation!
- Ethernet-based, OVS switches interconnected with virtual Ethernet pairs
- Topology-aware datagram forwarding and Shortest-Path Bridging
- Virtual compute nodes are connected directly to OVS switches over TAP interface
- One virtual compute node per virtual switch
- MPI traffic, network boot, port forwarding (DHCP, SSH, FTP, HTTP)



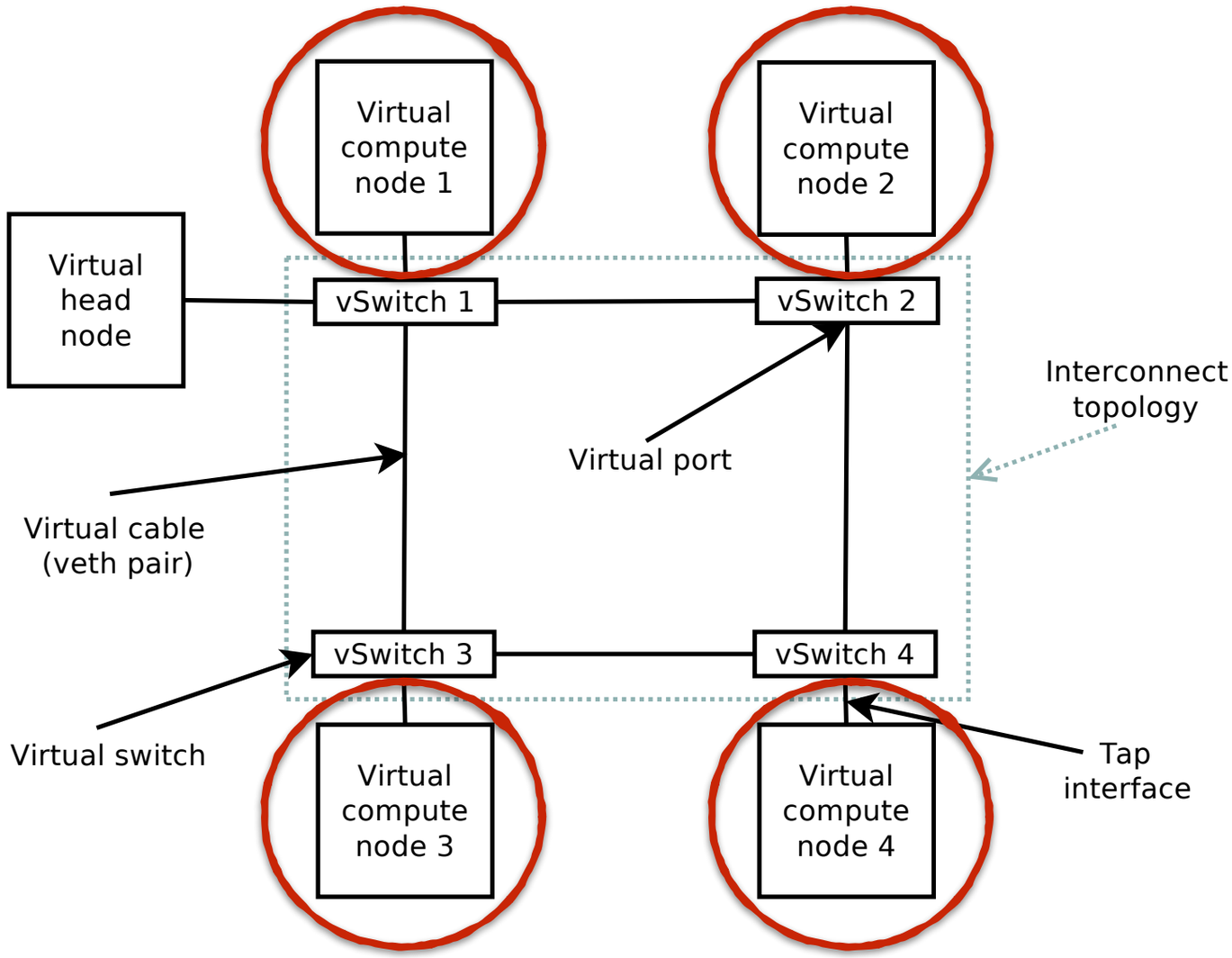
Virtual head node

- Login node to virtual head nodes
- Connects to the external world
- Provides DHCP server, to assign addresses to the virtual interconnect interfaces in virtual compute nodes
- Provides PXE server, to boot stateless virtual compute nodes
- The only node with a state, that stores configuration



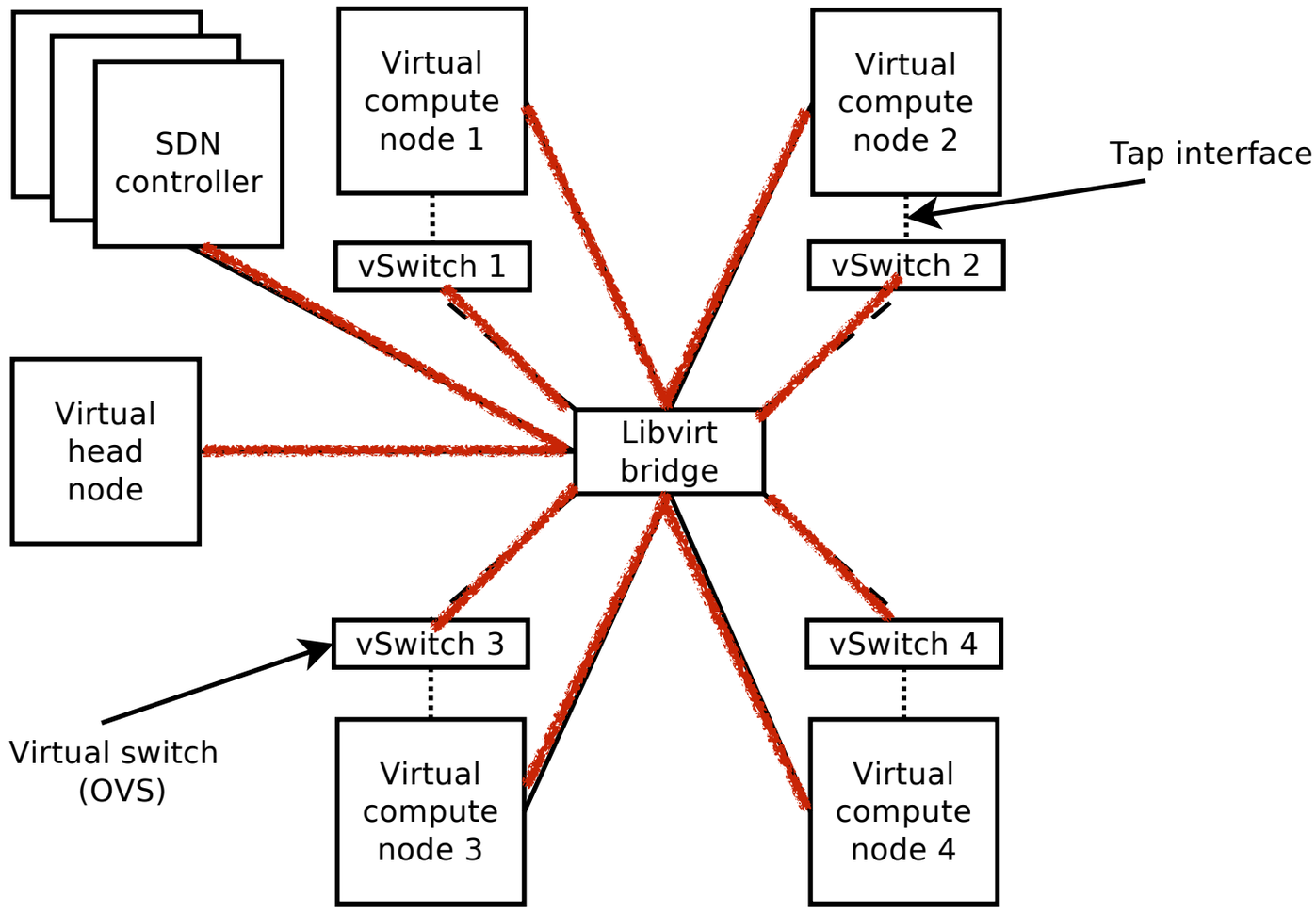
Virtual compute nodes

- Diskless and stateless
- Boot from head node
- Mount network attached storage over the management network
- Provide MPI run environment
- Communicate with each other over SSH over the virtual interconnect



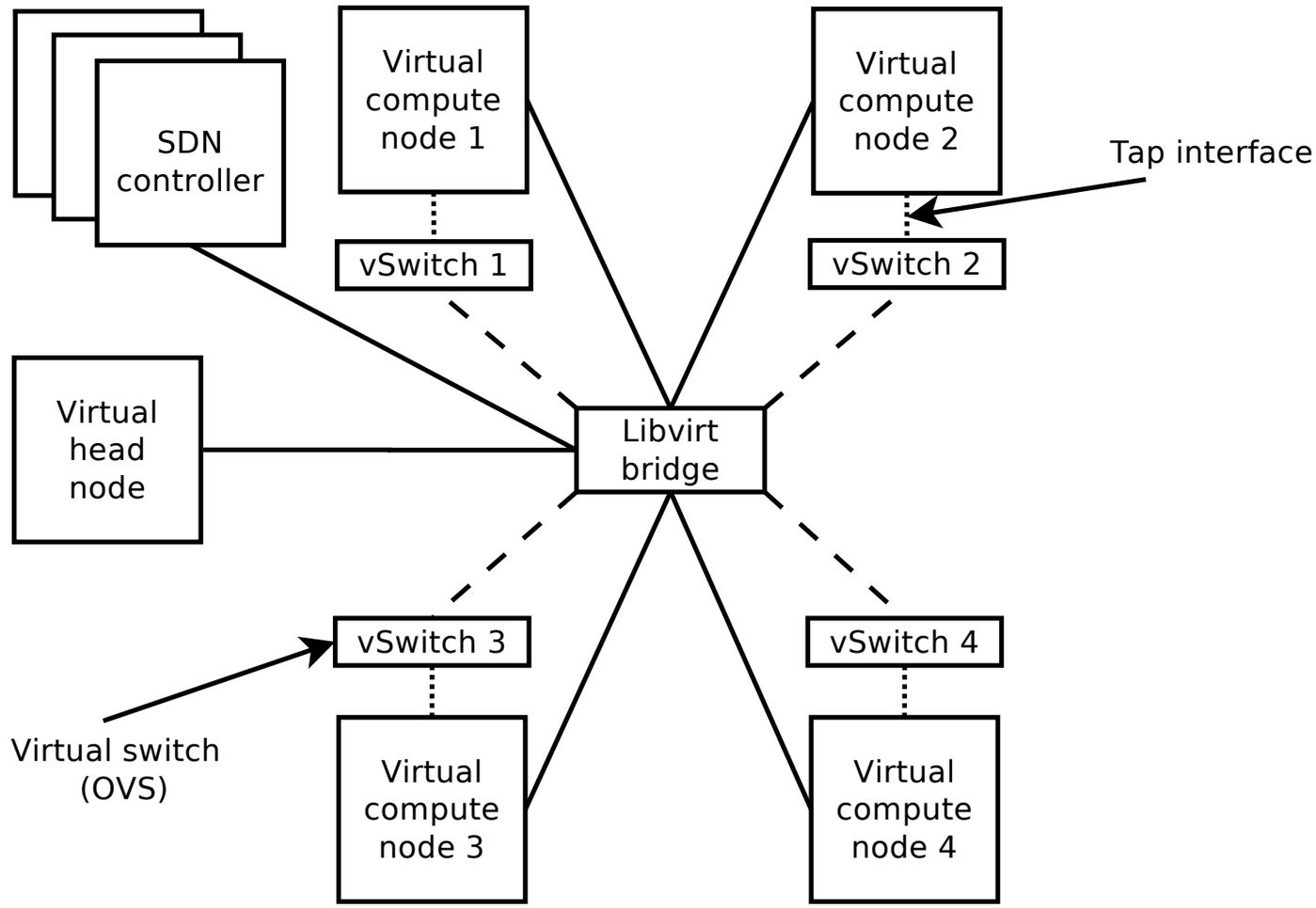
Management network

- Ethernet-based over libvirt switch
- Connects:
  - SDN controllers
  - Virtual compute nodes
  - Virtual head node
  - OVS switches
  - All possible additional components like traffic monitoring and response engine



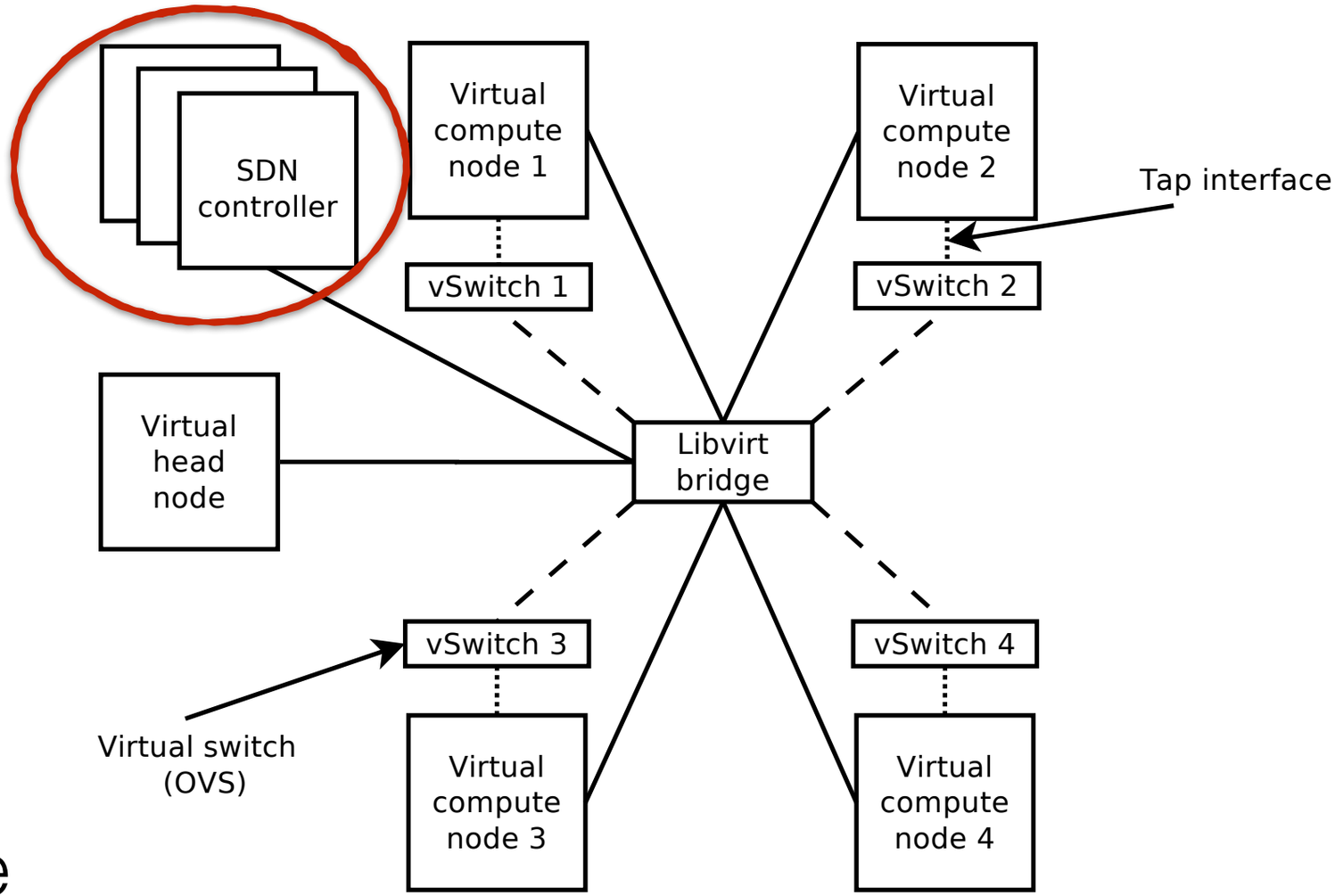
Management network

- OpenFlow traffic (SDN management protocol)
- QEMU Machine Protocol (VM management protocol)
- Network storage



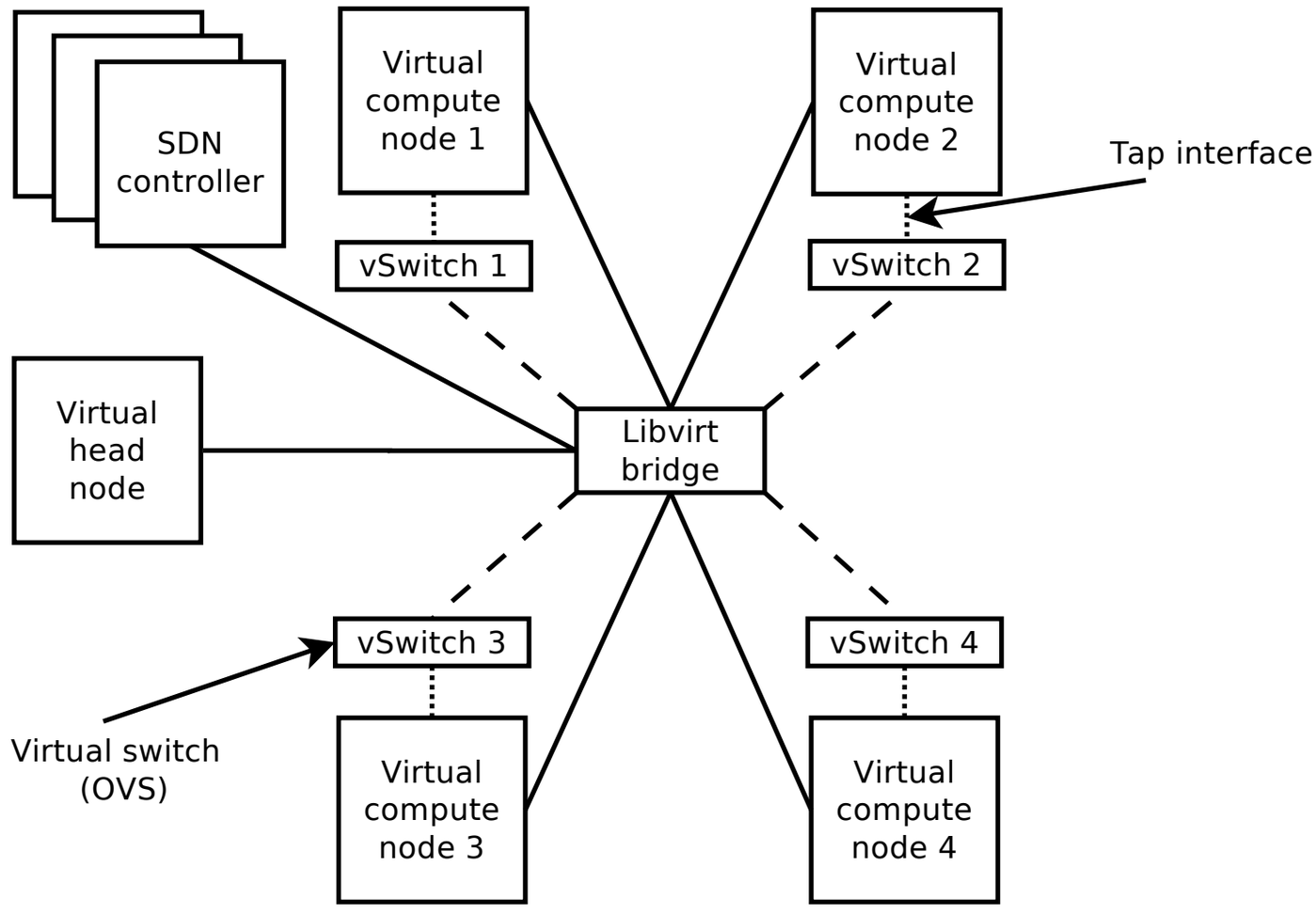
SDN controller nodes

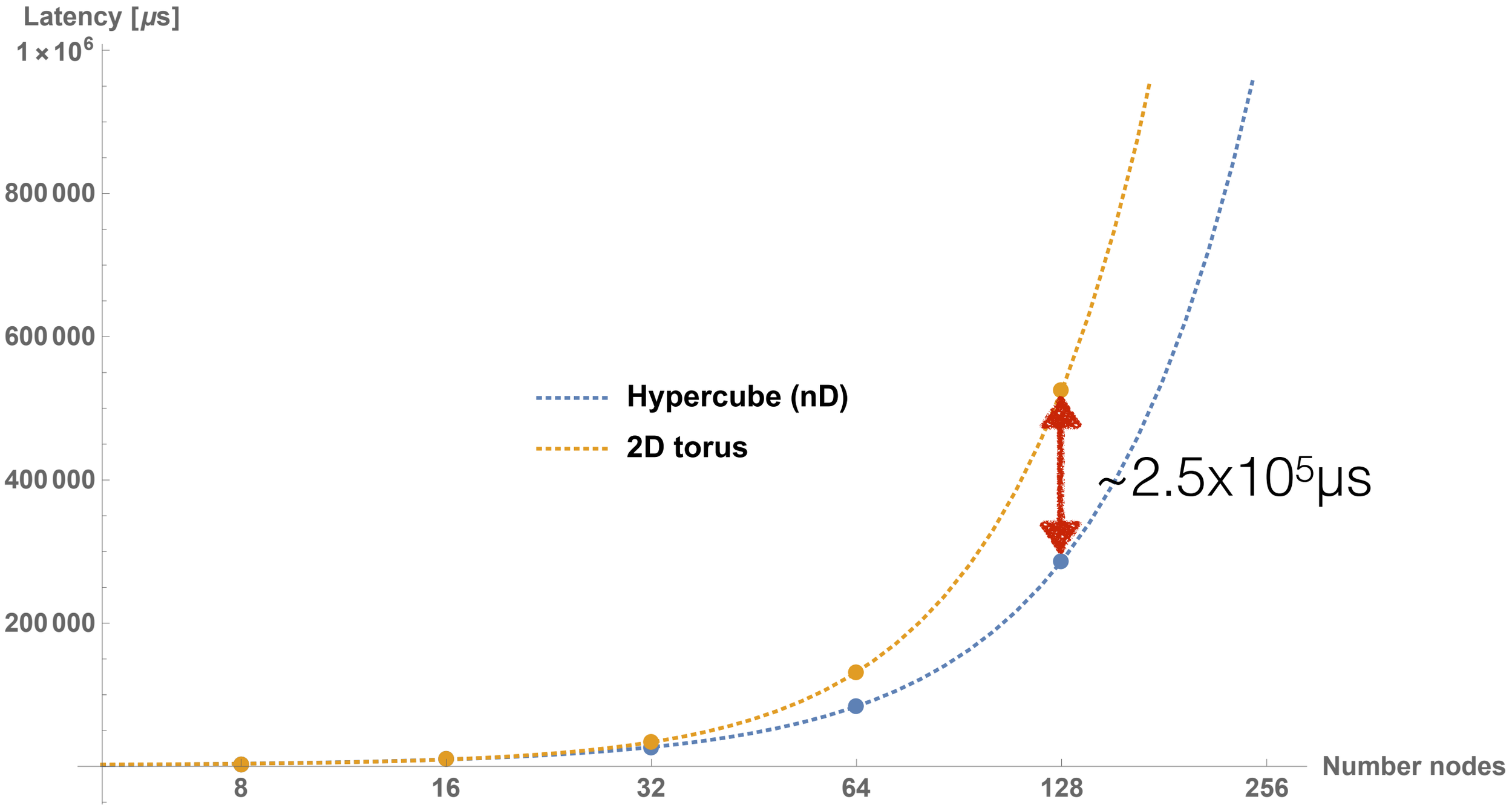
- Run ONOS SDN controller
- Diskless
- Boot from the head node
- Connects multiple instances of ONOS into a cluster over the management network



ONOS SDN controller

- Deployed as a single instance or a cluster of instances
- Connects to OVS switches over management network
- Topology aware with Shortest-Path Bridging
- Runs simple datagram forwarding application

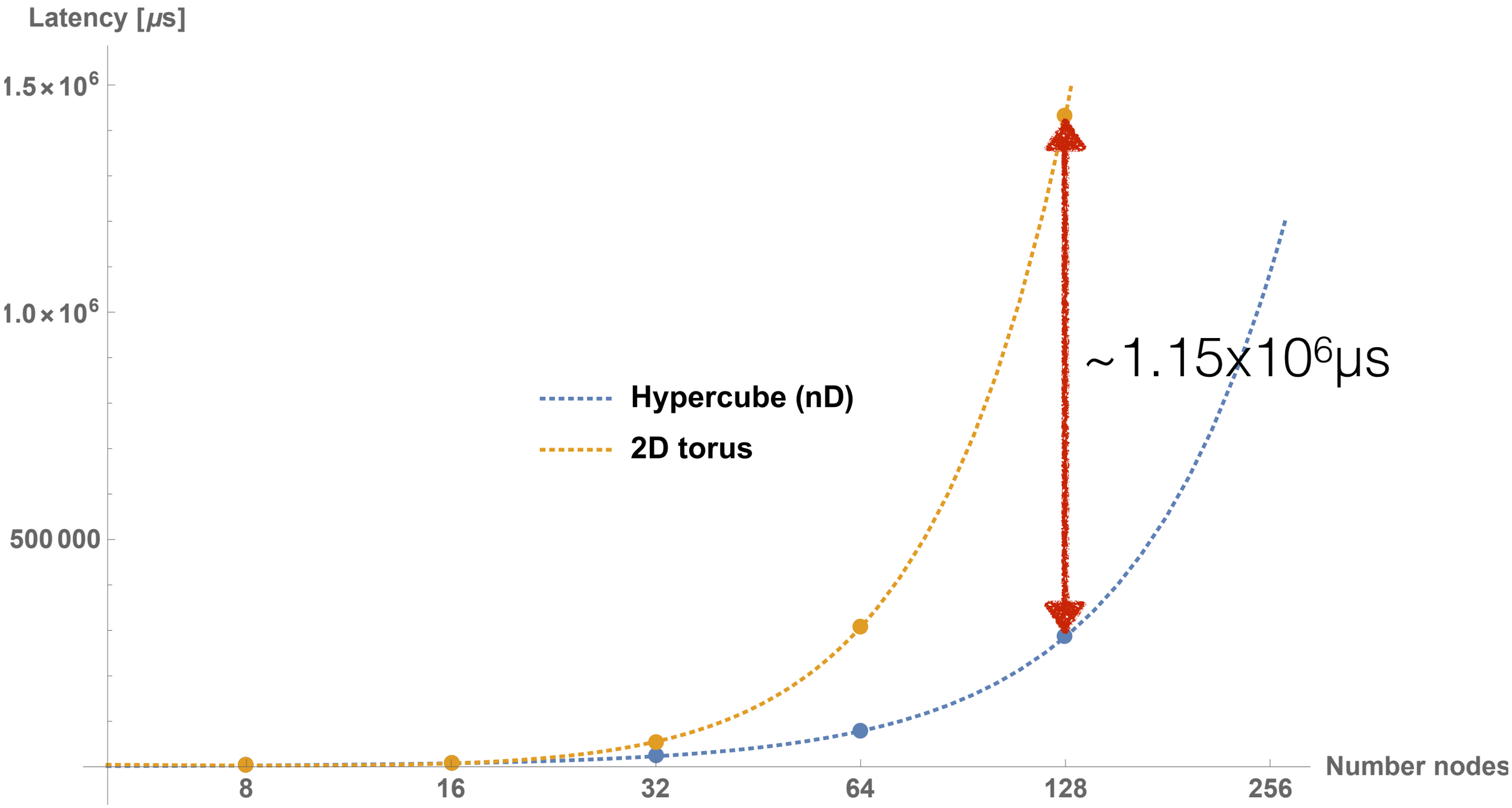




All to all OSU MPI benchmark with 256MB messages
Huawei system, 16 Intel(R) Xeon(R) CPU E7-8860 (128 cores @ 2.20 GHz) and 1,5TB memory

Computational Resource Centre
A*STAR

Stony Brook University



All gather OSU MPI benchmark with 256MB messages
Huawei system, 16 x Intel(R) Xeon(R) CPU E7-8860 (128 cores @ 2.20 GHz) and 1,5TB memory

Acknowledgements