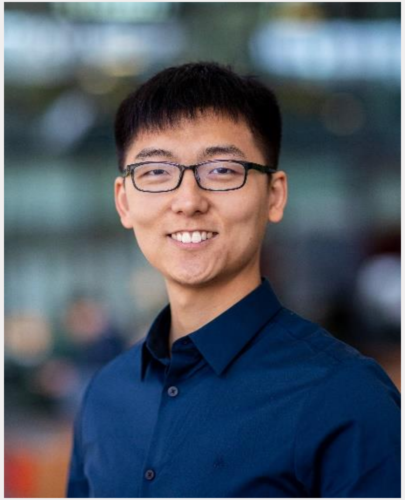
FOSDA: A Hybrid Disaggregated HPC Architecture based on Distributed Nanoseconds Optical Switches

Xiaotao Guo, Xuwei Xue, Bitao Pan, Fulong Yan, Georgios Exarchakos, Nicola Calabretta

Presenter: Xiaotao Guo
Eindhoven University of Technology
Email: x.guo@tue.nl

Presenter Biography



Xiaotao Guo is a Ph.D. candidate in Institute of Photonic Integration at Eindhoven University of Technology. His research interests include data center network, Software Defined Networking, and resource allocation algorithm.

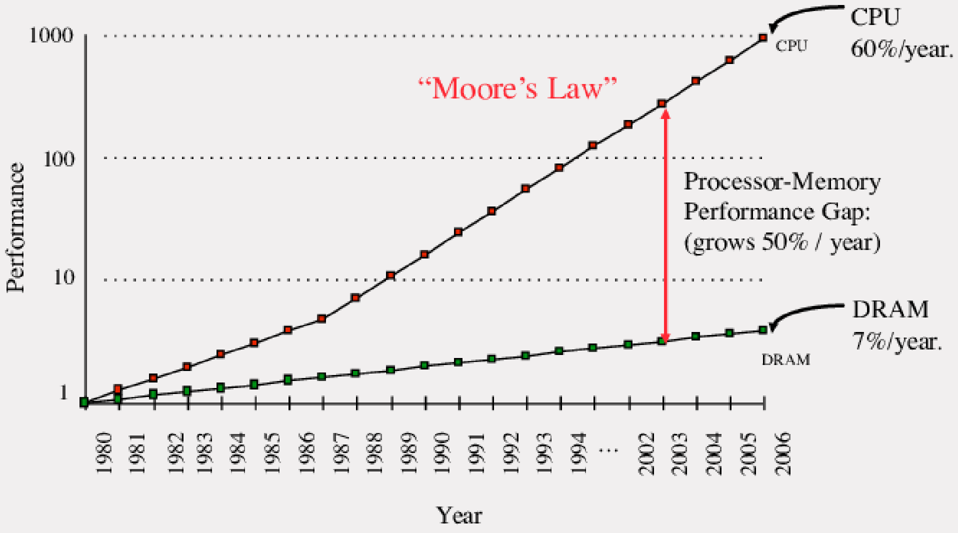
TU/e

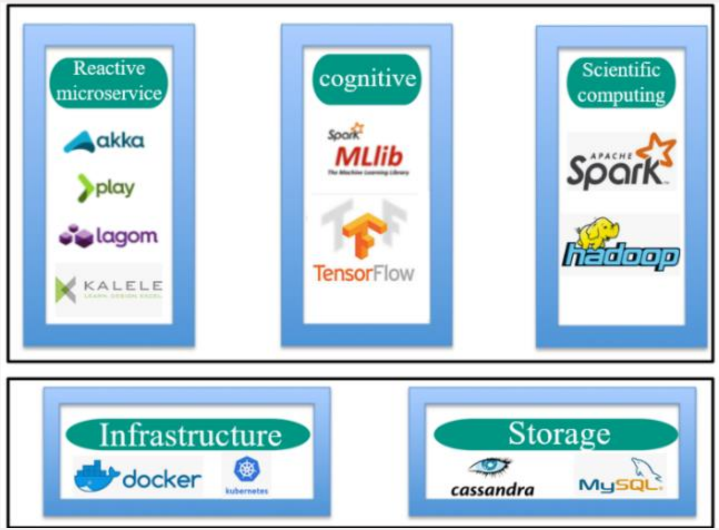
Outline

- Why Disaggregation in High Performance Computing (HPC) Network?
- Nanoseconds Optical Switch based Disaggregated Architecture
- Simulation Setup
    - Configurations of Disaggregated and Node-centric Architectures
    - Traffic Statistics from Two Node-centric HPC Networks
- Results
- Conclusions

Issues in Current HPC Network– "Performance Wall"

- Fixed amounts of hardware resources within the mainboard of computing node
- Continuous growing gap between CPU and memory performance
- Diverse workloads with even 4 orders of magnitude on memory over CPU demand.

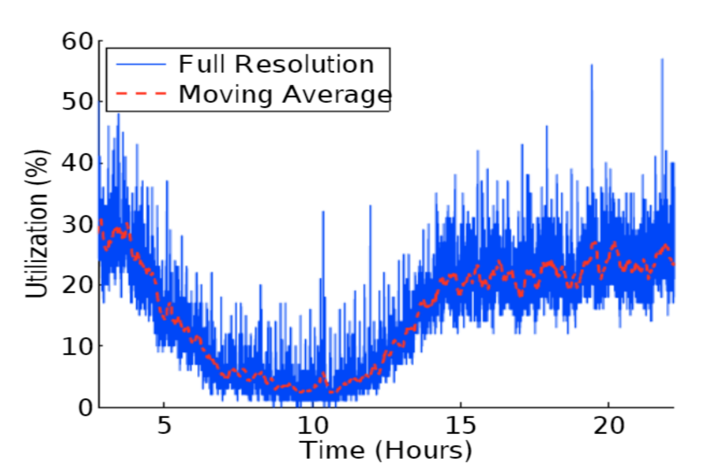


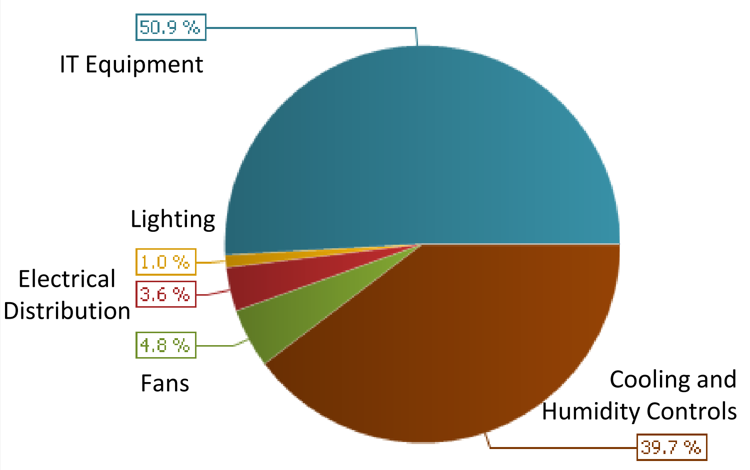


Issues in Current HPC Network– Resource, CAPEX and Energy Waste

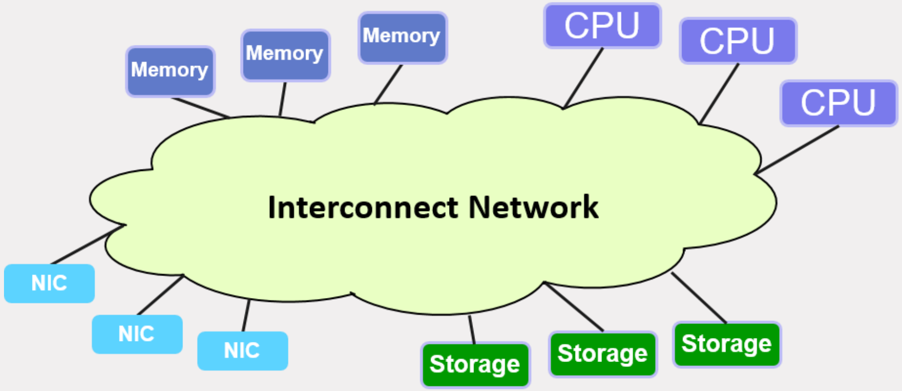
Mismatch between fixed hardware resource also results in:

- **Underutilized resources (even lower than 40%)**
- **Huge CAPEX waste since computing nodes account for 85% of total capital cost**
- **Underutilized resource takes up more than 50% energy consumption**





Promising Solution: Disaggregated HPC Network



Interconnect network requirements:

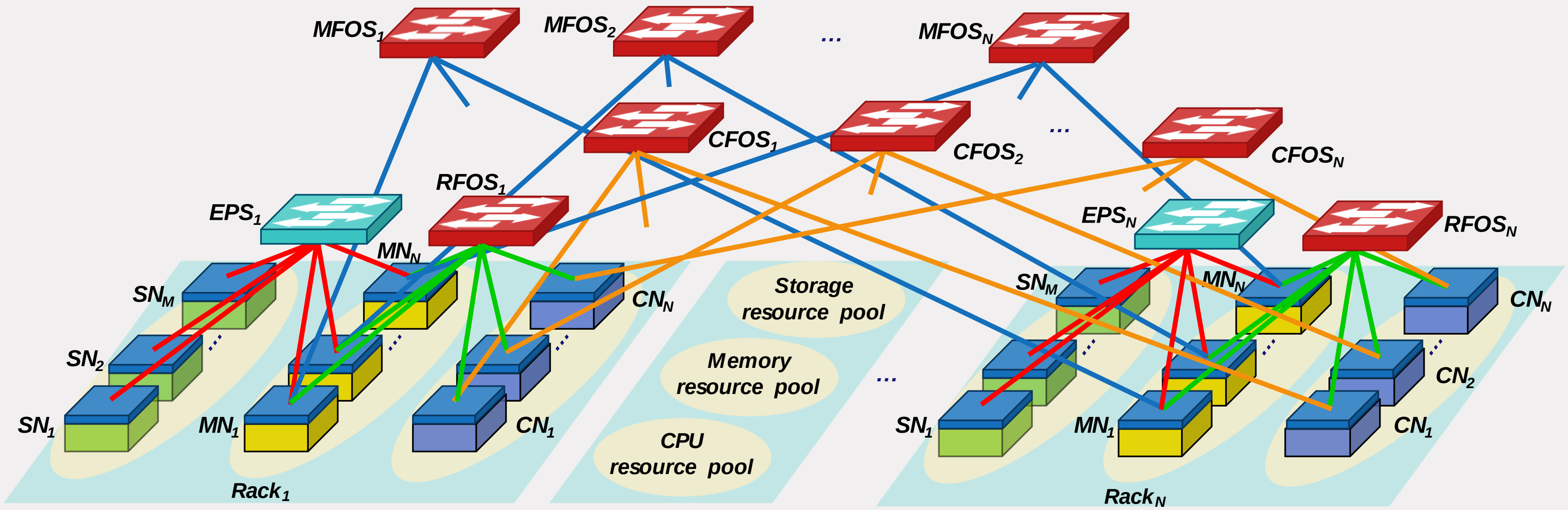
- **Fast transmission speed**
- **High bandwidth**
- **Low latency**
- **Scalability**

Different approaches [1, 2, 3]:

- Rack Scale Design (RSD): independent storage management system (coupled CPU and memory)
- A remote memory paging system (multi-layer electrical network may degrade performance)
- “dReDBox” network based on hybrid optical circuit and electrical switches (long switching time)

[1] Intel, “Intel Rack Scale Architecture Overview”, 2016.
[3] J. Gu, “Efficient memory disaggregation with infiniswap,” 2017.
[4] M. Bielski, “dReDBox: Materializing a full-stack rack-scale system prototype of a next-generation disaggregated datacenter,” 2018

FOSDA: Nanoseconds Optical Switch based Disaggregated Architecture



Properties:

- Fast switch speed (nanoseconds)
- Low latency for transmission
- High bandwidth capacity
- High scalability

[1] X. Guo, "RDON: a rack-scale disaggregated data center network based on a distributed fast optical switch," JOCN, 2020.

Simulation Setup

HPC2N:
- request rate: 17.44
- 120 nodes
- 240 cores
- 120GB memory
- 3 SCI network
- torus topology of 4×5×6

↕

FOSDA(12 racks):
- request rate: 17.44
- up to 144 nodes
- 240 cores
- 120GB memory
- splitting ratio F is 4
- TRX per node is 3

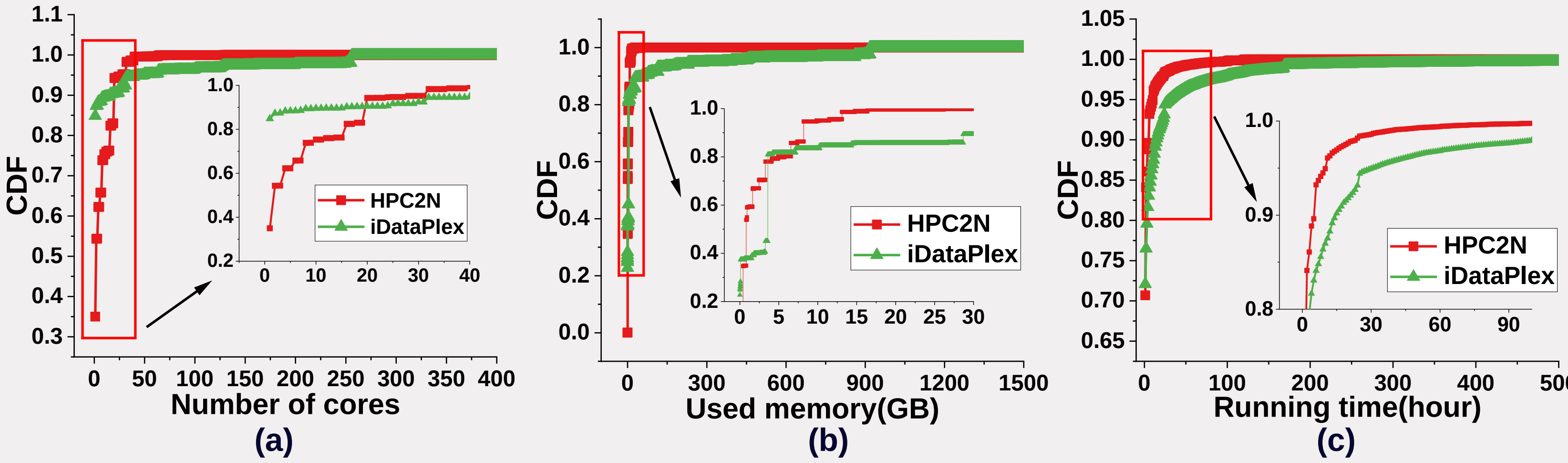
iDataPlex:
- request rate: 26.46
- 320 nodes
- 2560 cores
- 10240GB memory
- FDR InfiniBande network

↕

FOSDA(18 racks):
- request rate: 26.46
- up to 324 nodes
- 2560 cores
- 10240GB memory
- splitting ratio F is 6
- TRX per node is 3

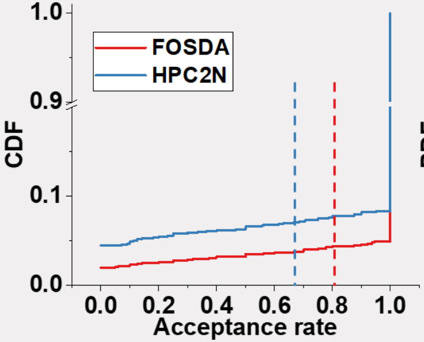
Components	**Specifications**		
	Type	***Power (W)***	***Cost ($)***
AMD Athlon MP2000+ processor	Idle Max	115 161	149
Intel Xeon E5-2660	Idle Max	116.4 194	1329
Memory	1G 32G 96G	0.373 11.85 35.55	6.5 209 637
NIC	Wulfkit3 10Gb/s 40Gb/s 56Gb/s	14 7 10.6 11.2	180 102 338 415
Transceiver	10Gb/s 40Gb/s 56Gb/s	1 3.5 4	18 59 84
Disk	HDD	6	154
Mellanox SX6536 Switch	648ports	9073	62,125
EPS	---	2/port	20/port
FOS	12ports 18ports 48ports	77 126 489	1140 2509 17612

Traffic Traces from Two Benchmark Node-centric Networks

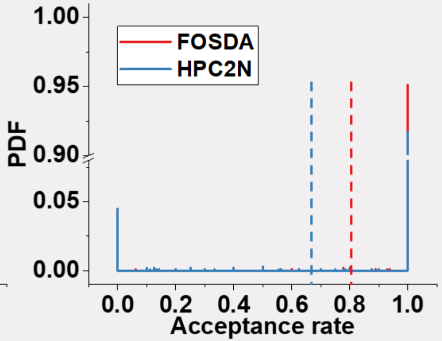


- Over 90% workloads have a CPU requirement of less than 50 cores in both architectures.
- Memory demand in HPC2N mainly ranges from 0 and 17GB, while 8.5% workloads requires more than 100GB memory in iDataPlex.
- More than 60% workloads have a running time of less than 2 hours in two HPC networks.

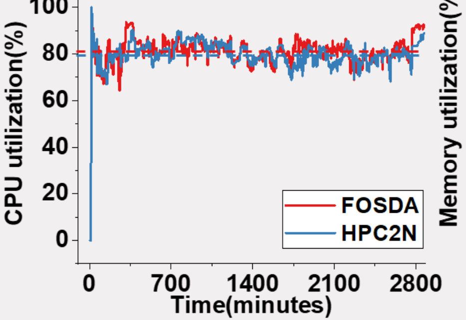
Comparison between FOSDA and HPC2N



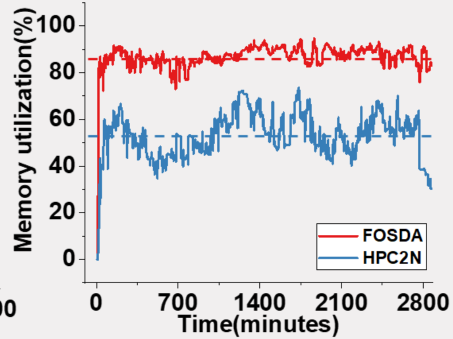
(a) CDF of acceptance rate



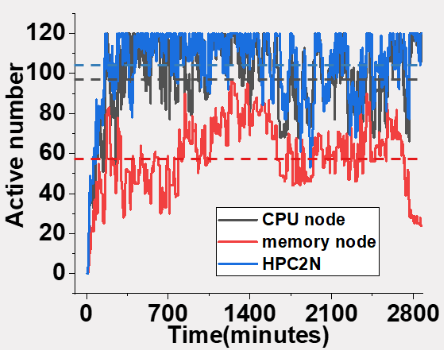
(b) PDF of acceptance rate



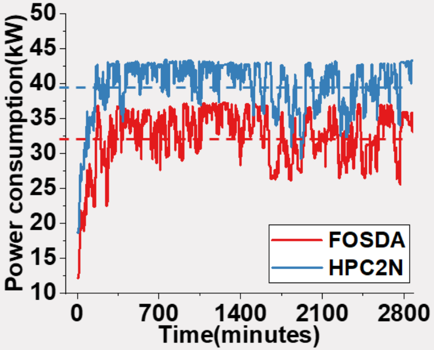
(c) CPU resource utilization



(d) Memory resource utilization



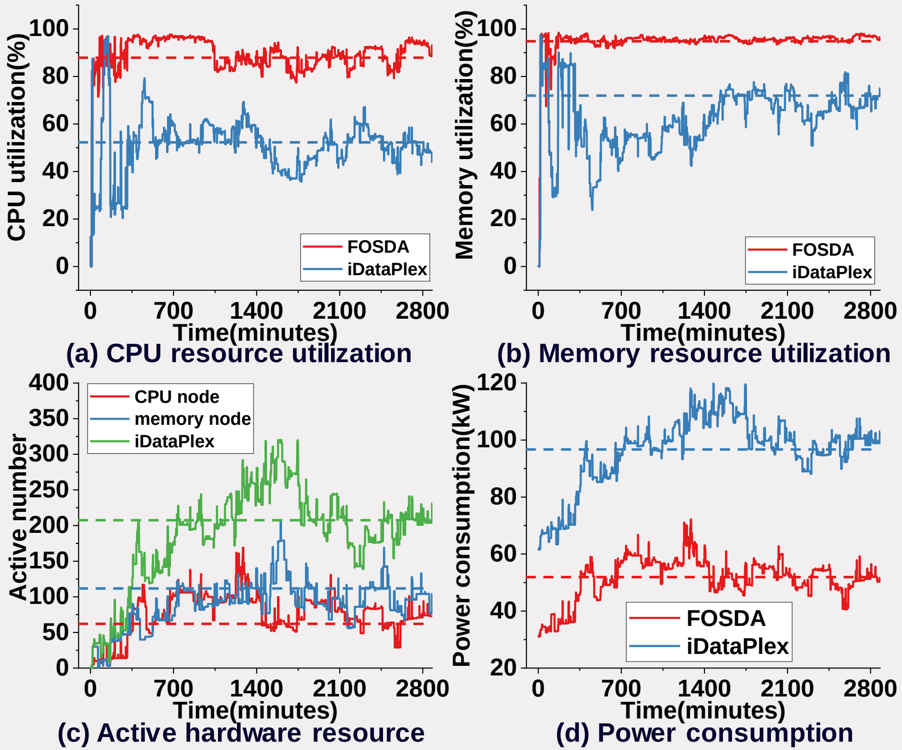
(e) Active hardware number



(f) Power consumption

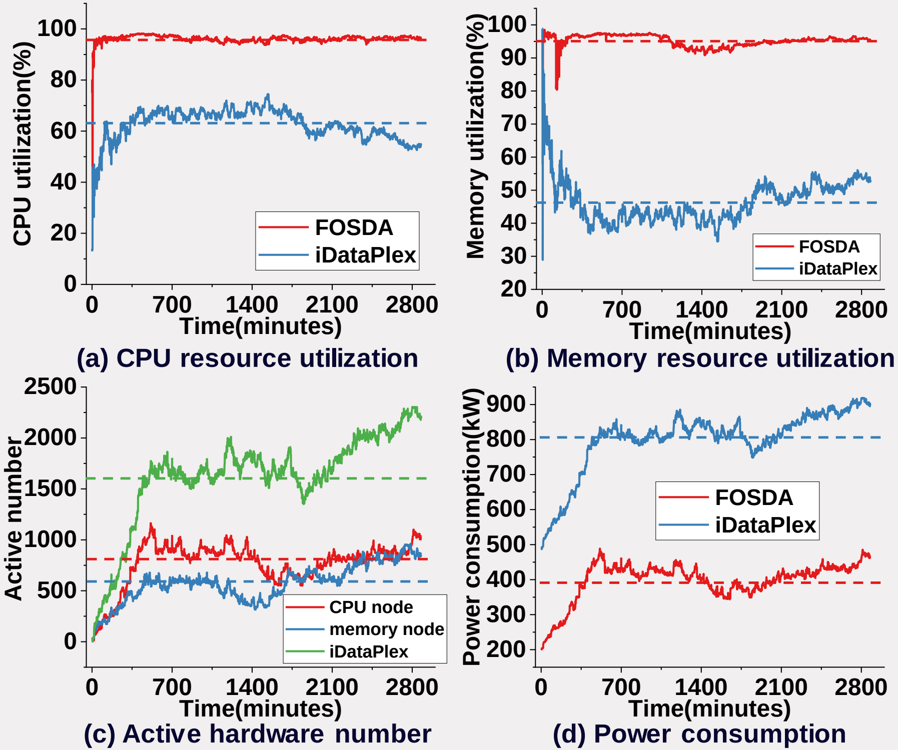
- FOSDA accepts 13% more workload requests
- FOSDA obtains 1.4% higher CPU and 33.4% higher memory utilization
- FOSDA saves 6.1% hardware and 18.7% power

Comparison between FOSDA and iDataPlex



- FOSDA obtains 36.6% higher CPU utilization
- FOSDA obtains 21.5% higher memory utilization
- FOSDA requires 45.5% less hardware
- FOSDA saves 46.8 % power consumption

Scalability of FOSDA (2304 nodes)



- **FOSDA obtains 33.6% higher CPU utilization**
- **FOSDA obtains 48.5% higher memory utilization**
- **FOSDA requires 52.5% less hardware**
- **FOSDA saves 50.4 % power consumption**

Capital and Operational cost Comparison

Architectures		Cost	
		Capital cost (k$)	*Operation Cost/year (k$)*
FOSDA	up to 144nodes up to 324nodes	346.8 1388.3	30.6 48.7
HPC2N	120 nodes	223.4	37.6
iDataPlex	320 nodes	1114	91.3

Compare with HPC2N:

- **FOSDA requires 35.6% higher capital cost**
- **FOSDA saves 18.6% operational cost**

Compare with iDataPlex:

- **FOSDA requires 19.8% higher capital cost**
- **FOSDA saves 46.7% operational cost**

Conclusion

- We present a novel disaggregated HPC architecture FOSDA based on distributed nanoseconds optical switches.
- Performance comparison of FOSDA and two benchmark node-centric HPC networks is based on realistic traffic traces.
- Compared with node-centric networks, FOSDA can accept up to 13% more workload requests, achieve up to 36.6% higher CPU and 21.5% higher memory utilizations with 45.5% less active hardware.
- In addition, FOSDA saves 46.8% power consumption compared with node-centric HPC network of 320 computing nodes.
- Moreover, compared with the node-centric HPC network, FOSDA requires 46.7% less operational cost with only 19.8% higher capital cost.

Thank You !