

LambdaRouter vs Big Fat Router

Charles Zhang, Venkat Vishwanath, Rajvikram Singh, Eric
He,
Luc Renambot, Jason Leigh

EVL Technical Document:
[20050312_zhang](#)

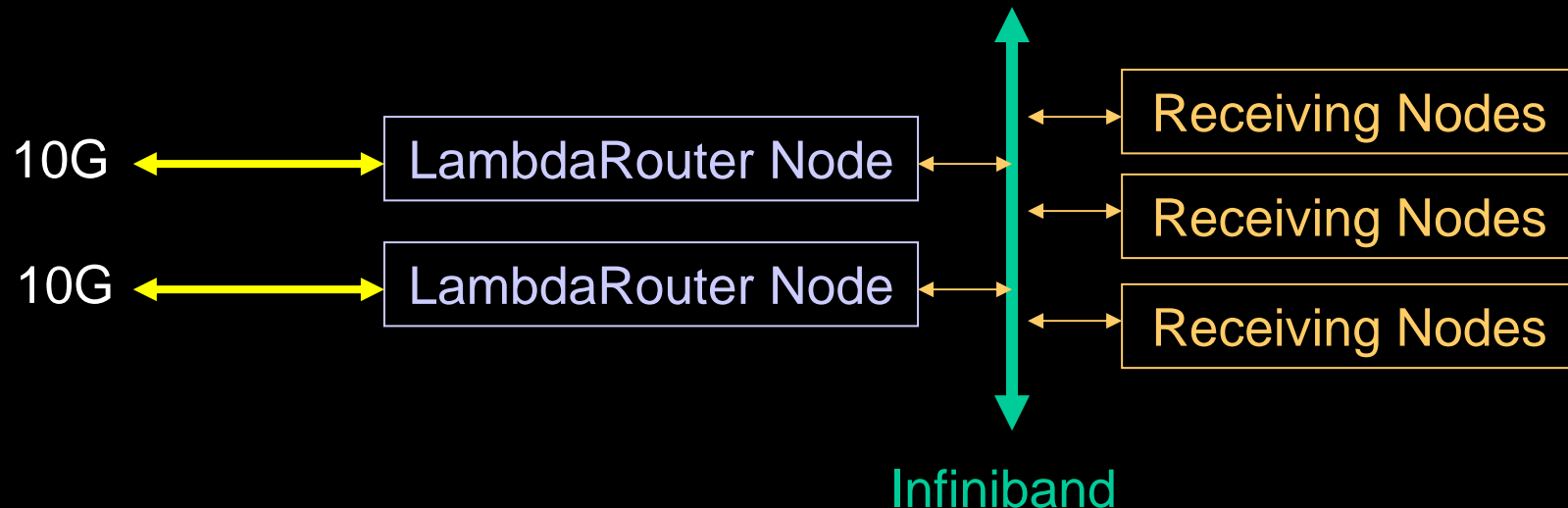
March 2005

Fundamental Question

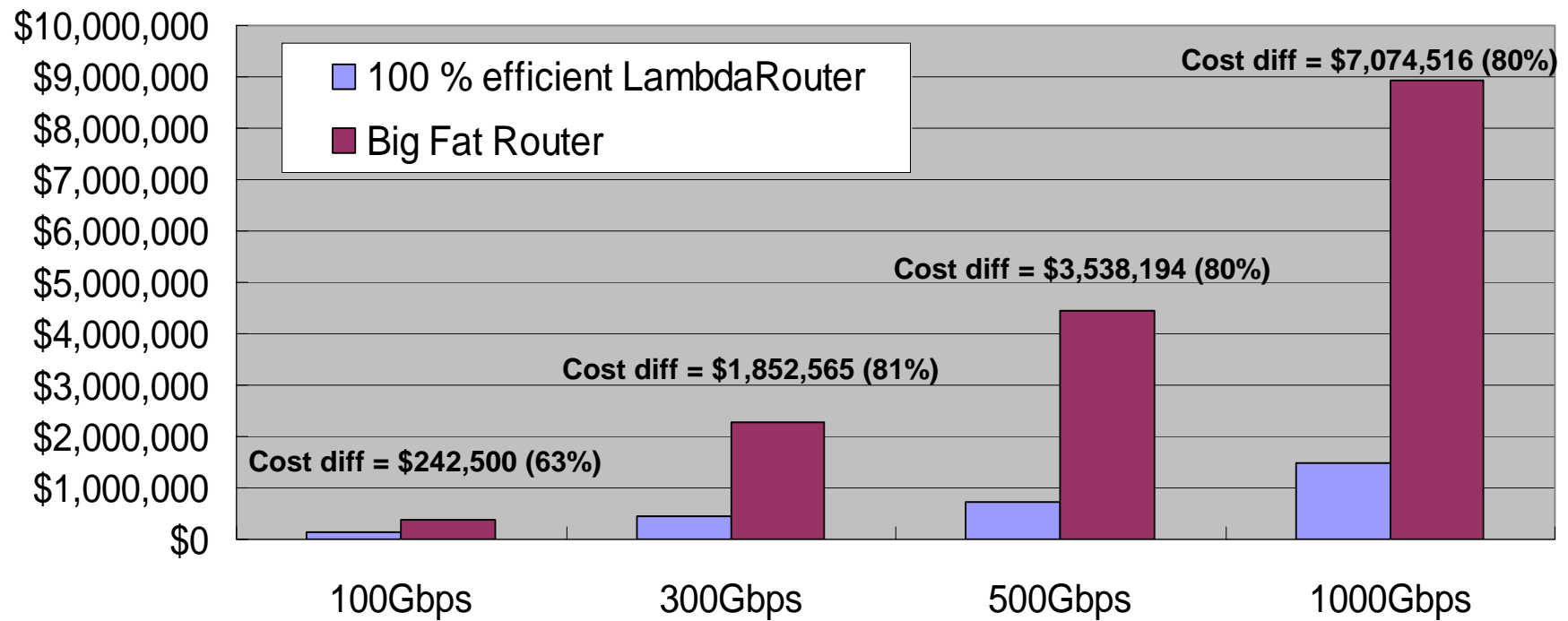
- How do you terminate tens and hundreds of gigabits of bandwidth at the end points so that cluster nodes at one end point can talk to arbitrary cluster nodes at the other end point?
- In Cees De Laat's case, how will he make StarPlane work?
- Buy a big fat router at every end point?
- Perhaps use end-point switching/routing.

What is end-point switching?

- Feed 10G links directly into fast PC nodes (LambdaRouter nodes) to “route/switch” data to final destination nodes via an infiniband / myrinet backplane.
- This is very similar to Lambda Optical System’s recent work with Hank Dardy to extend Infiniband over Lambdas.



Cost (\$) Comparison between LambdaRouter Architecture and Big Fat Router



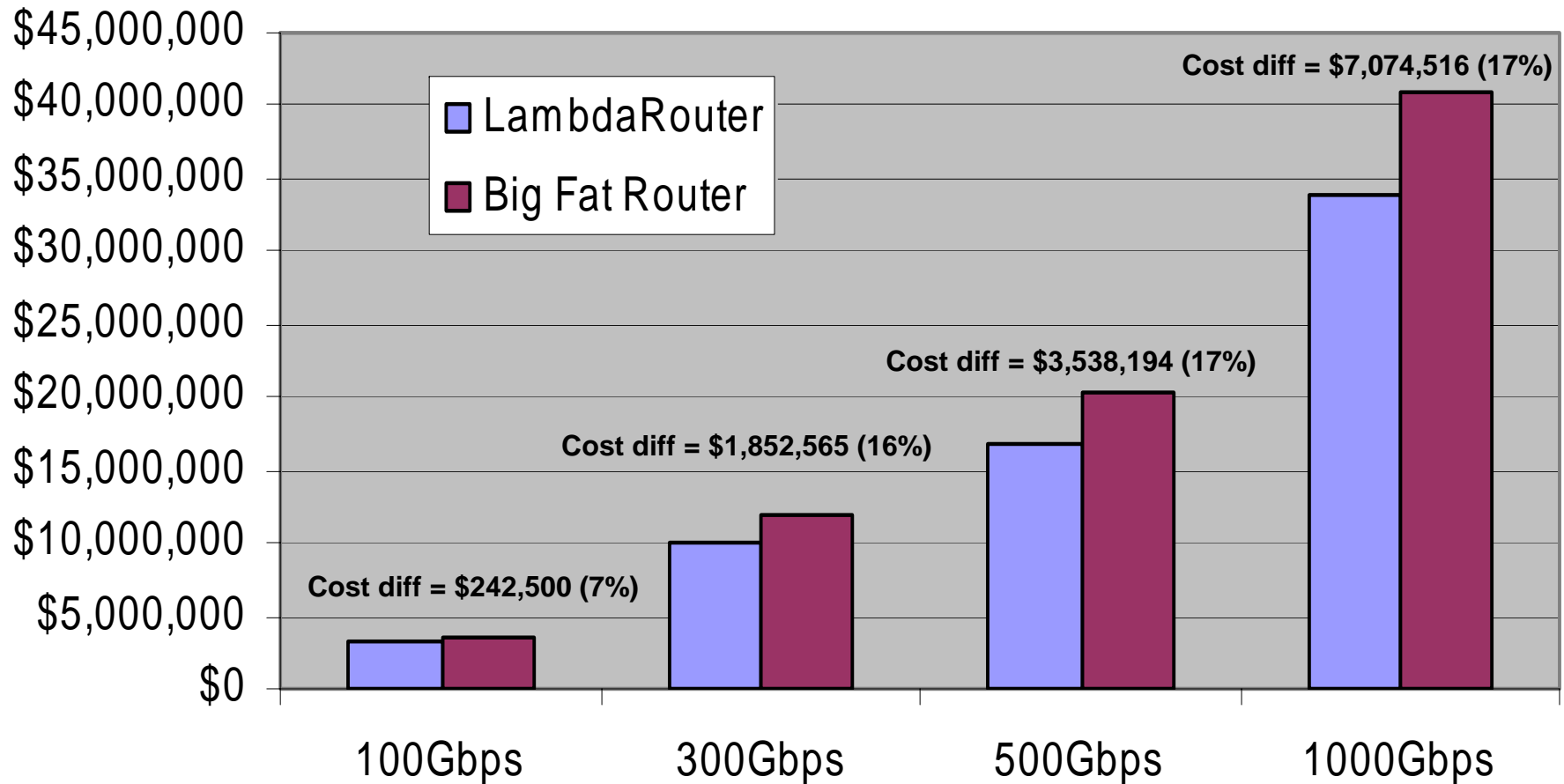
Explanation for previous graph

- The previous comparison assumes that the LambdaRouter is 100% efficient. I.e. it can perform as fast as a big fat router.
- We **do NOT** expect this to happen.
- The question is:
 - How efficient can we make the LambdaRouter?
 - Is the loss in efficiency acceptable considering the savings in \$?
- This is essentially like comparing commodity cluster computing against big Cray-like supercomputers.

What if we try to compensate for LambdaRouter inefficiency buying more Lambdas and wasting some of it.

- Assume a LambdaRouter is given a 10G link and it can route data at say 80% efficiency- ie it operates at 80% the rate of what a dedicated big fat router can perform.
- Assume we compensate for this loss of efficiency by spending more money to buy extra lambdas and LambdaRouter nodes so that the effective efficiency is back to 100%.
- The next graph shows what the total cost of the LambdaRouter solution would be if we had to factor in that extra incurred cost.

Cost (\$) Comparison between LambdaRouter Architecture and Big Fat Router assuming LambdaRouters could only achieve 80% efficiency but we compensated for it with more lambdas.



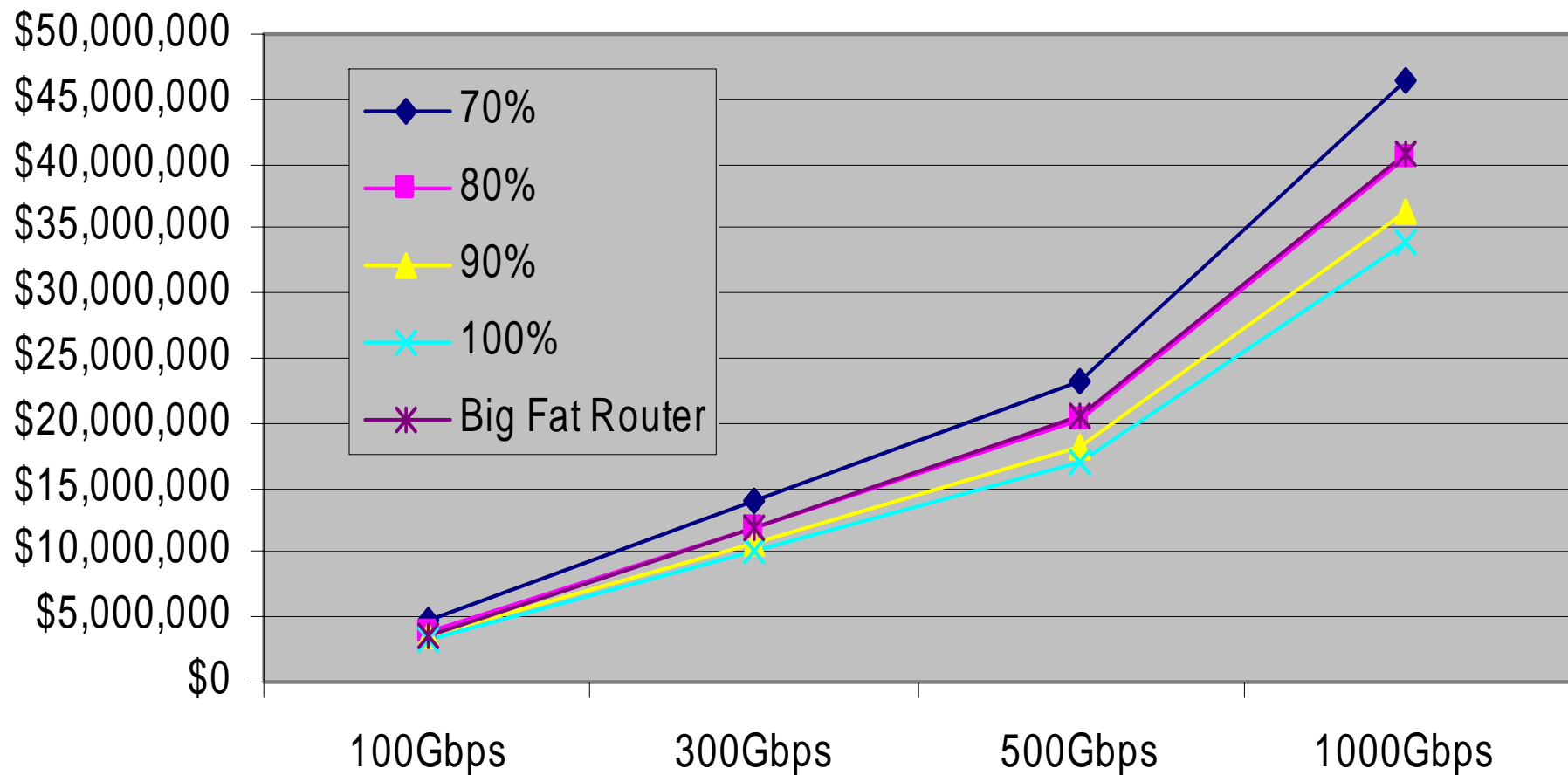
So What Does All This Imply?

- If you are willing to accept inefficiency in your routing solution then LambdaRouter is significantly cheaper as bandwidth increases.
- If you try to compensate for it by buying up more Lambdas to make up for the inefficiency then the cost difference between a big fat router diminishes.
- The major cost in the LambdaRouter solution are the Lambda and port costs.
- If that cost comes down it will bring the LambdaRouter solution's cost down.

Additional Notes

- What does NTT think about this?
 - Telcos are not interested because they don't expect end-users to be able to or want to deploy LambdaRouters.
- Do you really save \$?
 - You save \$ in hardware but you pay in middleware development costs.
- Looking at Foundry's new 10G switch offerings to see how much its cost compares against our models.
- Lambda Optical System's recent work with Hank Dardy to extend Infiniband over Lambdas is conceptually similar to this. Our LambdaRouter model allows us to analyze the cost benefits of LOS' system.

In case you are interested: Cost of LambdaRouter solution with compensation for different levels of inefficiency.



So How Did we Get Those Numbers?

[ie the gory details]

**Assumptions & Cost Model
(based on data gathered in early 2005)**

Disclaimer: Data are gathered from
vendor quotes where possible.

Observations from benchmarks

- We need a dual-Opteron (2.0 Ghz or better).
- If you want full utilization of 10G on a NIC we need TCP/UDP offloading. So we have to go with Chelsio, S2io etc. for the LambdaRouter nodes.
- InfiniBand has the capability to provide bi-directional 10Gbps bandwidth for each port.
- With current PCI-X 10GbE NIC, we can safely assume we can route 8Gbps per node.
 - When the PCI-Express 10 Gb cards are available, we can hope to route at full 10 Gbps.

Cost Model for LambdaRouter

- BackPlaneCostPerPort (around \$1,000)
- CostPerLambdaRouterNode (around \$12,000)
- efficiency (0.7, 0.8 and etc.)
- CostPerWideAreaLambda (around \$170,000 per lambda)
- CostPerDWDMPort (around \$150,000)
- BandwidthPerPC Gbps bandwidth allocated for each cluster node (in slide y=4);
- DesiredBW how much wide-area bandwidth you want to handle
- BandwidthPerLambda (assume 10Gb/s)

- **Total Cost = Cost of Backplane + Num Lambdas Needed * Cost of Lambdas and LambdaRouters and DWDM ports**
- **Cost of Backplane = BackPlaneCostPerPort * ((DesiredBW / BWPerPC) + Num of LambdaRouters Needed)**
- **Num of Lambdas Needed = DesiredBW / (BandwidthPerLambda * efficiency)**
- **Cost of Lambdas and LambdaRouters and DWDM ports = CostPerWideAreaLambda + CostPerLambdaRouterNode + costPerDWDMPort**
- **Num of LambdaRouters Needed** : each Lambda terminates in a LambdaRouter node so Num of Lambdas Needed = Num of LambdaRouters needed. Each LambdaRouter node also needs a port on the Infiniband backplane
- In a scenario where clusters are pipelined A->B->C some of the LR nodes in B will be allocated for A->B traffic and the rest will be allocated for B->C. This should not impose a significant bottleneck because the end cluster nodes still have a 10G NIC which has to receive and send data.

LambdaRouter 100Gbps uplink

24 Node Cluster with a 100Gbps routing infrastructure

$$\text{Total Cost} = \text{Cost of Backplane} + \text{Num Lambdas Needed} * \text{Cost of Lambdas and DWDM ports} + \text{Num Lambdas Needed} * \text{Cost of LambdaRouters}$$

Parts	Item #	Unit Price	Cost for different efficient systems			
			70%	80%	90%	100%
Performance Efficiency						
Lambda Nodes Needed			14	12	11	10
InfiniBand Switch (ISR-9024)	2	\$6,195.00	\$12,390.00	\$12,390.00	\$12,390.00	\$12,390.00
InfiniBand Adapter (HCA-400Ex)	24 + # of LRNodes	\$500.00	\$19,000.00	\$18,000.00	\$17,500.00	\$12,000.00
Cable (IC4-03)	40	\$100.00	\$4,000.00	\$4,000.00	\$4,000.00	\$4,000.00
SubTotal of Backplan			\$35,390.00	\$34,390.00	\$33,890.00	\$28,390.00
LambdaRouters		\$11,000.00	\$154,000.00	\$132,000.00	\$121,000.00	\$110,000.00
Lambdas and DWDM Ports		\$320,000.00	\$4,480,000.00	\$3,840,000.00	\$3,520,000.00	\$3,200,000.00
Total			\$4,669,390.00	\$4,006,390.00	\$3,674,890.00	\$3,338,390.00

$$\begin{aligned} \text{Total Cost of Big Fat Router} &= \text{Cost of NICs in each cluster node} + \text{Cost of Router} + \text{Cost of Lambdas \& DWDM ports} \\ &= \$72,000 + \$308,900 + \$3,200,000 = \$3,580,900 \end{aligned}$$

LambdaRouter 300Gbps uplink

72 Node System - 300Gbps uplink (actual 288Gbps)

$$\text{Total Cost} = \text{Cost of Backplane} + \text{Num Lambdas Needed} * \text{Cost of Lambdas and DWDM ports} + \text{Num Lambdas Needed} * \text{Cost of LambdaRouters}$$

Parts	Item #	Unit Price	Cost for different efficient systems			
			70%	80%	90%	100%
Performance Efficiency						
LambdaRouter Nodes Needed			42	36	32	30
InfiniBand Switch (ISR-9096)	1	\$12,967	\$12,967	\$12,967	\$12,967	\$12,967
InfiniBand Line Board (sLB-24)	4	\$7,117	\$28,468	\$28,468	\$28,468	\$28,468
InfiniBand Adapter (HCA-400Ex)	72 + # of LRNodes	\$500	\$57,000	\$54,000	\$52,000	\$51,000
Cable (IC4-03)	105	\$100	\$10,500	\$10,500	\$10,500	\$10,500
Subtotal of Backplane			\$108,935	\$105,935	\$103,935	\$102,935
LambdaRouters		\$11,000	\$462,000	\$396,000	\$352,000	\$330,000
Lambdas & DWDM Ports		\$320,000	\$13,440,000	\$11,520,000	\$10,240,000	\$9,600,000
Total			\$14,010,935	\$12,021,935	\$10,695,935	\$10,032,935

$$\begin{aligned} \text{Total Cost of Big Fat Router} &= \text{Cost of NICs in each cluster node} + \text{Cost of Router} + \text{Cost of Lambdas \& DWDM ports} \\ &= \$216,000 + \$2,069,500 + \$9,600,000 = \$11,885,500 \end{aligned}$$

LambdaRouter 500Gbps uplink

120 Node System - 500Gbps uplink (actual 480Gbps)

$$\text{Total Cost} = \text{Cost of Backplane} + \text{Num Lambdas Needed} * \text{Cost of Lambdas and DWDM ports} + \text{Num Lambdas Needed} * \text{Cost of LambdaRouters}$$

Parts	Item #	Unit Price	Cost for different efficient systems			
			70%	80%	90%	100%
Performance Efficiency			70%	80%	90%	100%
LambdaRouter Nodes Needed			69	60	54	50
InfiniBand Switch (ISR-9096)	1	\$19,467	\$19,467	\$19,467	\$19,467	\$19,467
InfiniBand Line Board (sLB-24)	8	\$7,117	\$56,936	\$56,936	\$56,936	\$56,936
InfiniBand Adapter (HCA-400Ex)	120 + # of LRNodes	\$500	\$94,500	\$90,000	\$87,000	\$85,000
Cable (IC4-03)	200	\$100	\$20,000	\$20,000	\$20,000	\$20,000
Subtotal of Backplane			\$190,903	\$186,403	\$183,403	\$181,403
LambdaRouters		\$11,000	\$759,000	\$660,000	\$594,000	\$550,000
Lambdas & DWDM Ports		\$320,000	\$22,080,000	\$19,200,000	\$17,280,000	\$16,000,000
Total			\$23,220,806	\$20,232,806	\$18,240,806	\$16,912,806

$$\begin{aligned} \text{Total Cost of Big Fat Router} &= \text{Cost of NICs in each cluster node} + \text{Cost of Router} + \text{Cost of Lambdas \& DWDM ports} \\ &= \$432,000 + \$4,019,000 + \$16,000,000 = \$20,451,000 \end{aligned}$$

LambdaRouter 1Tbps uplink

240 Node System – 1Tbps uplink (actual 960Gbps)

$$\text{Total Cost} = \text{Cost of Backplane} + \text{Num Lambdas Needed} * \text{Cost of Lambdas and DWDM ports} + \text{Num Lambdas Needed} * \text{Cost of LambdaRouters}$$

Parts	Item #	Unit Price	Cost for different efficient systems			
			70%	80%	90%	100%
Performance Efficiency			70%	80%	90%	100%
LambdaRouter Nodes Needed			138	120	107	100
InfiniBand Switch (ISR-9096)	2	\$19,467	\$38,934	\$38,934	\$38,934	\$38,934
InfiniBand Line Board (sLB-24)	24	\$7,117	\$170,808	\$170,808	\$170,808	\$170,808
InfiniBand Adapter (HCA-400Ex)	240 + # of LRNodes	\$500	\$189,000	\$180,000	\$173,500	\$120,000
Cable (IC4-03)	400	\$100	\$40,000	\$40,000	\$40,000	\$40,000
Subtotal of Backplane			\$438,742	\$429,742	\$423,242	\$369,742
LambdaRouters		\$11,000	\$1,518,000	\$1,320,000	\$1,177,000	\$1,100,000
Lambdas & DWDM Ports		\$320,000	\$44,160,000	\$38,400,000	\$34,240,000	\$32,000,000
Total			\$46,555,484	\$40,579,484	\$36,263,484	\$33,839,484

$$\begin{aligned} \text{Total Cost of Big Fat Router} &= \text{Cost of NICs in each cluster node} + \text{Cost of Router} + \text{Cost of Lambdas \& DWDM ports} \\ &= \$864,000 + \$8,050,000 + \$32,000,000 = \$40,914,000 \end{aligned}$$

Cost Model for Traditional Big Fat Router

- Total Cost = cost of NICs in each cluster node + Cost of Router + Cost of Lambdas + Cost of DWDM ports
- Cost of Lambdas = $(\text{DesiredBW} / \text{BandwidthPerLambda}) * \text{CostPerWideAreaLambda}$
- Next slide has quotes from vendors (their names are withheld) on how much it would cost to build a 100 Gbps, 300 Gbps, 500 Gbps and 1 Tbps switch.

Cost based on Vendor's Quotes

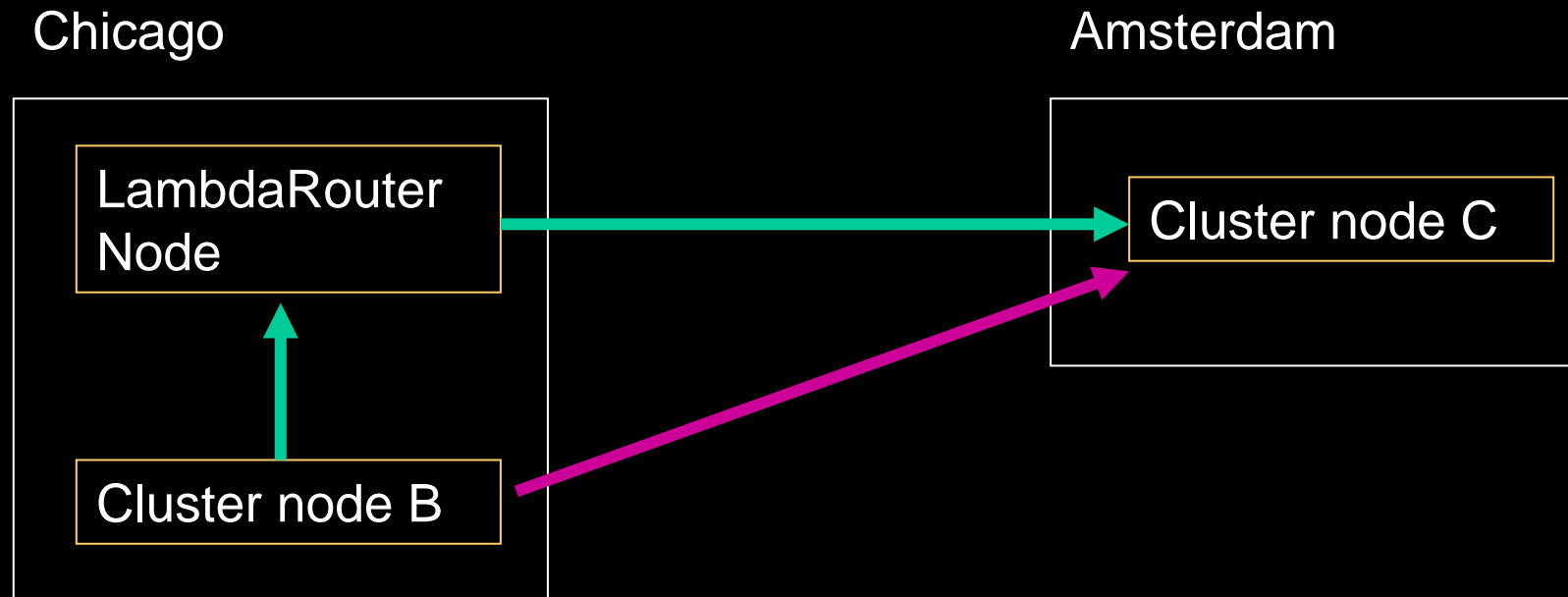
(Vendor's name withheld)

100Gbps uplink router	\$308,900
300Gbps uplink router	\$2,069,500
500Gbps uplink router	\$4,019,000
1Tbps uplink router	\$8,050,000

So What about Latency in the LambdaRouter?

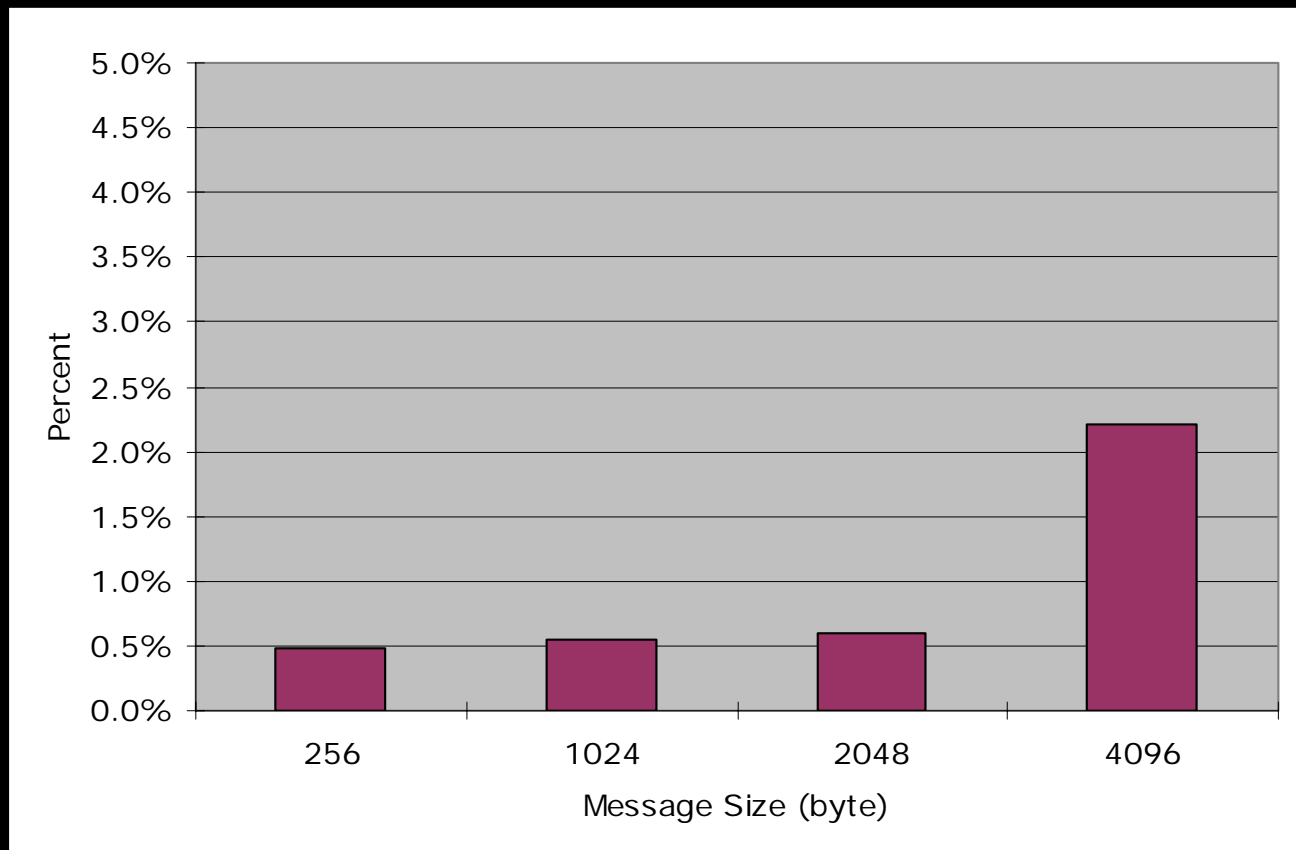
- Conducted a simple latency test routing various sized packets from Chicago to Amsterdam.
- Disclaimer: these results are only preliminary. There were done to get a sense of what the lower bound on latency might be in a real situation.

Latency Testbed



What is the difference in latency between the green path and the red path?

Latency overhead by “routing” through an intermediate LambdaRouter node



Latency over the trans-oceanic communication (from Chicago to Amsterdam) is around 55 milliseconds

The experiment was done on 1Gbe link.