

Network Planning for Rich Media Content and Application Providers

by Jeffrey Papen, Peak Web Consulting

As the Internet economy experiences a major resurgence, utilization is soaring and new applications are impacting every aspect of consumers' lives, from dating to social networking to entertainment.

This paper is intended to help the network manager or director-level staff member to understand better the challenges related to sudden and substantial growth. The goal is to educate readers about general network best practices for large-content sites, to provide a primer on what questions should be asked of your potential providers, and to point out technology and industry pitfalls that network planners need to be aware of.

A New Internet Economy Presents New Network Challenges

As the Internet economy experiences a major resurgence, utilization is soaring and new applications are impacting every aspect of consumers' lives, from dating to social networking to entertainment. As a result, companies once again need to invest in network infrastructures. But while companies are expanding and attempting to hire, existing staffs and resources cannot keep pace. Web 2.0 companies – and others who are supporting event-driven traffic, providing highly interactive offerings, or delivering rich-media content such as video – are faced with new issues. They need to make informed decisions about creating network infrastructures that will support and adapt to expansion and spikes in demand into the long term.

In particular, companies need to:

1. Carefully select a primary transit provider, and learn when multihoming makes sense
2. Effectively evaluate colocation options
3. Efficiently interconnect datacenters that reside both in the same metro area and in multiple regions
4. When appropriate, learn how to take advantage of network peering.

Large or small, emerging or established, you need to make the right network decisions so that the exponential growth that your company is hoping for is growth that you – and your network partners – are prepared to handle.

Consider the Issues Across All of Your Network Components

TRANSIT NETWORKS – NOT ALL TRANSIT IS CREATED EQUAL

When it comes to transit, you need sufficient capacity, including redundant failover capacity, to reach your customers regardless of their Internet Service Providers (ISPs). If your network needs to send 4 to 6 Gigabits to a single ISP to



Once you have a solid relationship with a Tier 1 provider that is able to scale quickly and that delivers consistently reliable traffic across its own network, you can work to find alternate routes to a subset of your customers using other carriers.

support a customer event, you need to be sure that your primary transit provider can get the job done. A surprisingly high number of “Tier 1” provider networks have not been keeping up with the changing times or growing to meet continued demand. These providers, in a large number of occurrences, have connectivity between their networks, but do not have any appreciable growth capacity. You may learn too late, if you choose the wrong provider, that the network is out of capacity.

There’s a reason that transit providers come in at very different price points. It can make economical sense, after you have a primary transit provider, and assuming that you have the staff and the desire, to bring down your blended per-megabit rate by multihoming to less expensive, Tier 2 ISPs. Once you have a solid relationship with a Tier 1 provider that is able to scale quickly and that delivers consistently reliable traffic across its own network, you can work to find alternate routes to a subset of your customers using other carriers. Multihoming in this manner must be done carefully, and with a thorough understanding of both your traffic destination volumes and of each ISP’s specific strengths and weaknesses.

Keep the following questions in mind as you search for a primary transit provider:

- *Does the provider rely heavily on other providers?* How much traffic does your transit provider send to settlement-free direct peers, or to upstream ISPs? A surprisingly small number of ISPs exchange all of their traffic with other ISPs on a settlement-free basis. Many ISPs that call themselves Tier 1 buy transit from other ISPs, many of which are foreign. Heavy reliance on these providers can be a liability, since the Service Level Agreements (SLAs) they offer cannot cover your traffic once it leaves their network. Also beware of limits to the amount of traffic that you can burst to your customers.
- *How much capacity does the provider have to your target networks?* You need to understand where your traffic is going. And you need to have a relationship with an ISP that can give you accurate answers about how much capacity is available on its network, if they have capacity at all.
- *How well connected is the transit provider to your customers’ networks?* If you know what eyeball networks are the primary consumers of your site, you can ask a potential transit provider how they are connected. For example, all of the large cable-modem networks, with one exception, are directly connected to the Level 3 Network. You may be surprised to learn that many other “Tier 1” ISPs are one or even two networks away from your customers. Be aware that with every network “hop” between you and your customers, you add an order of magnitude to the number of things that can go wrong.
- *Does the provider have failover capabilities that you can rely on?* Do you have a primary transit provider that can effectively handle failover and reliably deliver your traffic, even when things go wrong? Every provider experiences an occasional fiber cut or equipment failure. Even if your network is multi-homed, your primary ISP needs to be able to handle your entire site’s traffic load if another ISP fails. Look for a provider that is willing to deliver periodic capacity information to your customer networks. You don’t want to rely on a sales staff to learn whether your provider can meet your requirements; you may find out during an outage that you’re the victim of marketing fluff.
- *Does the provider have capacity to support growth across new target networks?* The glut of capacity that you could count on several years ago is no longer

standard. Over the past several years, fewer providers have been able to invest in their networks to make consistent upgrades to accommodate the Internet's growth. Don't find yourself in a position where you have to work with your ISP's engineers to cherry-pick the networks for which they have capacity. And don't align yourself with a provider that will sell you circuits that can't be installed immediately, or that has to delay your orders while extra capacity is being built into your market. These are clear indicators of poor capacity planning and provisioning.

- *Can the provider turn up a 10 Gigabit Ethernet (10 GigE) interface quickly?* Any company anticipating substantial growth needs to rely on a core provider that can turn up a large amount of fully burstable capacity in under two weeks. A 10 GigE port should be business-as-usual to a quality provider – you do not want to be their first install. Look for a provider that can quickly turn up a 10 GigE interface without having to juggle other customer orders to accommodate a quick-turn event – and without outrageous non-recurring charge (NRC) fees.

MULTIHOMING – IMPROVE YOUR PERFORMANCE AND LOWER YOUR COSTS BY ESTABLISHING RELATIONSHIPS WITH MULTIPLE ISPS

The goals of multihoming are to reduce your reliance on a single ISP and to drive down your blended per-megabit rate. But you don't want to find yourself in a situation where you're sending traffic to low-cost providers and just hoping that it works. Successful multihoming involves establishing a tight relationship with a high-quality ISP. With careful planning, you can send portions of your traffic to Tier 2 ISPs without impacting your customers' experience.

But what kind of expertise does this require? Your staff needs to have thorough knowledge of how much traffic you're sending to different target networks. From there, they need relationships with engineers at various ISPs that can give you honest answers about what will work well and which traffic you're better off routing through an alternate provider. Your staff also needs to have the ability to manage effective Border Gateway Protocol (BGP) policy routing, including multi-level failover policies, and the expertise to implement and maintain economic-based routing policies.

There are “BGP Special Sauce” appliances maturing, from vendors such as NetVMG or Route Science, which can help with multihoming issues. However, the appliances can be expensive and there are significant limitations to how much traffic they can accurately manage.

COLOCATION – LEARN TO COMPARE APPLES TO APPLES – THEN LEARN TO SNIFF OUT A LEMON

With different colocation vendors touting their amounts of space, their cooling per square foot or their price for electricity, comparing services can be confusing. Make an “apples-to-apples” comparison by calculating the cost in dollars per “fully loaded” kilowatt. First, figure out the total power draw that the facility can cool (do not figure the power built out, because no circuit is drawn at 100 percent). Then determine the total cost for cooling, space, and power. Once you have the total amount of power your company can utilize, divide by the total bill to get a “fully loaded” cost. This way, you can calculate the unit cost in a way that translates different datacenter specifications and billing methodologies.

When searching for colocation space, expect that you are going to buy space that you're not going to use. Most datacenters available today were built to

Successful multihoming involves establishing a tight relationship with a high-quality ISP. With careful planning, you can send portions of your traffic to Tier 2 ISPs without impacting your customers' experience.

When searching for colocation space, expect that you are going to buy space that you're not going to use. Most datacenters available today were built to handle between 80 and 125 watts per square foot.

handle between 80 and 125 watts per square foot. Over time, colocation users are installing more and more densely packed servers, to the point that it is not uncommon to see 15 or more kilowatts of power in a single rack. In existing datacenters, this single rack would require between 120 and 200 square feet of space to cool properly. It is an unfortunate fact of life that most colocation facilities will have a large amount of unused space, but you should expect any colocation provider worth its salt not to allow over three 20 amp 120 volt circuits per rack. A provider that allows more power is allowing a situation in which an outage is waiting to happen. Additionally, it is often the case that putting fewer servers per rack will provide better cooling and lower operational costs. Servers will ultimately run better, because they are able to maintain a proper operating environment and they are easier to maintain by operations staff when not tightly packed in a small environment.

Another key to making an informed colocation decision is learning what facilities it takes to cool equipment. Cooling densities are very low compared to the heat generated by a typical rack. You could possibly need 300 square feet to cool a single rack. Beware of vendors who claim that they can cool 200 or 300 watts per square foot. In this case, marketing claims need to take a back seat to the laws of physics: any vendor must have over three feet of raised floor to cool close to 200 watts per square foot. If the facility is situated on a cement slab with ambient cooling, the limits are around 125 watts per square foot. And beware of vendors that claim they can cool "whatever you need," because the costs to retrofit or add cooling capacity can be outrageous.

METRO CONNECTIVITY

Don't be afraid of the dark

As legacy datacenters run out of available power and cooling, large content providers find their servers spread across multiple datacenters within a given metro area. One of the biggest challenges for growing these networks is finding a scalable, reliable and cost-effective way to interconnect these multiple datacenters. Depending on the distances between your colocation facilities, multiple options exist for interconnecting, each with their own pros and cons.

Metro Dark Fiber

Dark fiber is simply a very long Ethernet cable that uses different optics to drive the light farther. Metro dark fiber offers the highest reliability, because it has the fewest moving parts that can fail or that require maintenance. The only thing that could take down a dark fiber network is a backhoe, which becomes a non-issue when you have a redundant (i.e. ring) configuration. With metro dark fiber, you can also run 1 Gbps, 10 Gbps, or Dense Wavelength Division Multiplexing (DWDM) on the same fiber for the same fixed rate.

Using DWDM systems in conjunction with dark fiber to multiplex multiple interfaces across the same physical fiber yields the highest performance for cost. DWDM systems available today can drive 320 Gbps around a metro dark fiber ring. In this scenario, dark fiber is approximately 38 times less expensive than lit services (assuming \$10,000 MRC for the dark fiber ring versus 32 times \$12,000 MRC for 32 10 Gbps lit services). However, dark fiber is limited by distance; typically 80 fiber kilometers is the maximum distance, depending on the optics you choose.

Metro Lit Fiber

If your datacenters are more than 80 fiber kilometers apart, then you will most likely need lit services. Unlike dark fiber, metro lit fiber can travel to your distant locations, but always at a higher price. In most cases, if you have to purchase service over 10 Gbps pipes, you buy your circuits in 10 Gbps chunks even if your traffic doesn't grow in such large increments. Look for a provider that will sell rate-limited 10 Gbps circuits. If distance between datacenters makes lit fiber the only option for you, find a vendor that will dedicate to your network both the metro Ethernet fiber and the equipment to light the fiber. Avoid connecting with a company that will oversubscribe Ethernet circuits or put you on the same circuit with other customers. In these scenarios, a problem with another customer's migration, installation or upgrade could mean downtime for you.

Even with dedicated circuits, there are still more "moving parts" to be aware of in lit gear, so take care to compare vendors' maintenance windows and SLAs. Based on my years of experience, I've found that a lit service is about three orders of magnitude more likely to have an outage than a dark service. Lit services can either be protected or unprotected. If possible, have the vendor provide two handoffs in each location that run live-live, leaving the failover to you. In this way, you run two 10 Gbps links at 20 Gbps, which is diverse but not protected, rather than only one 10 Gbps link of diverse and protected service. The basic premise is to avoid paying for back-up service that you only use during emergencies.

If distance between datacenters makes lit fiber the only option for you, find a vendor that will dedicate to your network both the metro Ethernet fiber and the equipment to light the fiber. Avoid connecting with a company that will oversubscribe Ethernet circuits or put you on the same circuit with other customers.

INTERCONNECTING THE WIDE AREA NETWORK: WAN CIRCUITS

Know what you're paying for — and how it's protected

As companies expand across the country for both disaster-recovery and performance reasons, interconnecting disparate datacenters with WAN services presents a new challenge as well as options for an array of services and technologies. The technology for WAN circuits is very similar to metro lit services, but the naming and redundancy options can be quite different. Comparing WAN services means understanding the differences between unprotected Wavelength versus protected SONET, or private-line services, and understanding LAN PHY versus WAN PHY handoffs.

LAN PHY versus WAN PHY services

LAN PHY offers the ability to use 10 GigE interfaces, and WAN PHY requires use of 10 Gbps OC-192 interfaces. The important thing to keep in mind is that some vendors may charge differently between LAN PHY and WAN PHY. With some switches, you can change the encapsulation to either LAN PHY or WAN PHY using software, so the same Gigabit Interface Connector (GBIC) works for either.

Wavelength versus SONET services

Both Wavelength and SONET services use the same Optical Carrier (OC) encapsulation, as opposed to Ethernet encapsulation. A wavelength is a lit point-to-point service that does not have built-in failover protection, whereas SONET usually has automatic 50-millisecond re-route protection. The OC encapsulation on a wavelength service does not imply the protection received from SONET service. SONET's circuit-level protection means that SONET can be approximately three times as expensive as a wavelength service. To maintain the proper level of redundancy using wavelength services, you must have topology-level redundancy, which is typically found in ring configurations. When properly

When considering any network service, be very careful about collapsed laterals, which happen when two sides of a ring follow the same physical path. When this happens, a backhoe could cut both the primary and backup path on your ring with one physical swipe.

configured, your traffic can be rerouted if the wavelength's path is cut. When cost is a primary consideration, consider wavelength services. If, during an outage, your application cannot handle the latency of a sub-optimal re-route path, then consider the more expensive SONET services. In either case, do not buy services from a vendor that cannot provide detailed fiber maps showing exactly how their WAN paths traverse the country and enter or exit any of your datacenters. Wavelength and SONET services can be converted to an Ethernet hand-off to the customer, however, the cost increase for the service usually far outweighs the difference in cost for the customer router interfaces.

When considering any network service, be very careful about collapsed laterals, which happen when two sides of a ring follow the same physical path. When this happens, a backhoe could cut both the primary and backup path on your ring with one physical swipe. Seek providers that have redundant ingresses and egresses into your colocation site, and who can, under the terms of a non-disclosure agreement (NDA), provide detailed fiber maps. In some datacenters, there may be no way to avoid a collapsed lateral getting out to the street. In these situations, ensure that fiber takes a diverse path once it exits the building. Redundant ingresses into a building can cost tens of thousands of dollars per foot to create, which typically make them cost-prohibitive.

PEERING – THE BENEFITS ARE MEASURABLE, BUT BREAKING IN TAKES EXPERTISE

The Holy Grail for most large content companies is to send as much traffic as possible directly to other networks “for free.” When you establish a peering relationship, you exchange traffic directly between networks with the exchange of little or no money. But because peering efforts are motivated by both economic and political factors, it is not always as easy as it may appear – as one would hope – to establish relationships with ISPs.

Realizing the benefits of peering in both performance and cost savings requires a solid understanding of peering policies and extensive expertise in running national backbones using traffic-shaping techniques and multi-tier auto-rerouting BGP policies. Who are your largest target ISPs? Where do you have to be physically to interconnect with them? How many time zones do you have to be in? What ISPs can you realistically peer with? For which ISPs will you meet public versus private peering criteria? What are appropriate traffic ratios and how will your network maintain them, both on a per-link and an intercity basis? How do you failover properly? When is hot-potato routing an alternative to cold-potato routing? How should you handle peers that are only in a single location? How do you build a number of BGP sessions safely, in a way that helps you protect your routing table from being polluted if a peer leaks inappropriate information? How do you decide how big your backbone should be? If you're pushing 20 Gbps, does that mean you need a 10 Gbps backbone, or can you get by with OC-48, GigE or Fast Ethernet? When is it appropriate or possible to establish settlement-based versus settlement-free peering? At what rate is transit less expensive than a peering network, which would make peering moot?

In addition to learning how to physically manage peering, you need to understand how to initiate peering agreements. The peering community is very tight and extremely communicative about which networks are more trouble than they are worth as peers. Breaking into this community can be a challenge. Peering

relationships always boil down to layers 8 and 9 of the Open Systems Interconnection (OSI) model – the economic and political layers. They have nothing to do with the quality of connectivity or capacity between networks, with making customers happy, or with improving the end-user experience. You need to approach each of your potential peers with an understanding of the economics and politics that motivate them to best position your network as an attractive peering candidate.

Summary

As you're looking to implement and manage complex networks that involve multiple markets, a wide range of bandwidth requirements, and a range of network technologies, you need a core provider that has the capabilities to provide a comprehensive range of services. When you can count on your provider to deliver everything from colocation to transit to inter-colo connectivity, you're linked to the foundation you need to grow, whether you're among the smallest or the largest content delivery networks. A high-quality provider can help you keep up with rapid growth, while effectively reducing your risk of outages. Once you've established these core services, you may supplement your network with Tier 2 providers and/or peer with other networks to reduce your costs. But this must be done with experience, expertise and a thorough understanding of all of the dangers and caveats involved.

To learn more networking issues faced by rich media content providers and Web 2.0 companies, or to learn more about Peak Web Consulting, please e-mail Jeffrey Papen at jeffrey@peakwebconsulting.com.

When you can count on your provider to deliver everything from colocation to transit to inter-colo connectivity, you're linked to the foundation you need to grow, whether you're among the smallest or the largest content delivery networks.

Jeffrey Papen, Founder, Peak Web Consulting. JNCIE #116, Certified Juniper Instructor

Jeffrey Papen implemented the BGP load-balancing and multihoming policies at both Yahoo! and Excite@Home from 1998 - 2003. Mr. Papen's responsibilities at Yahoo! included developing the BGP multihoming policy to optimize network performance and balance traffic levels to over 200 peering and transit ISPs, while meeting transit commitments within 1 percent of their theoretical minimums, and still maintaining the highest possible performance. Mr. Papen has also developed numerous network analysis tools including: SQL-based bandwidth usage and billing reconciliation (happydog), ISP SLA testing (Glacier), BGP transit analysis (TUNDRA), and non-invasive multihoming performance testing (Alpine). Since leaving Yahoo! in March of 2003, Mr. Papen has operated Peak Web Consulting whose team of engineers provide networking services to some of the largest Web 2.0 companies.

PEAK
WEB CONSULTING