Ovum
TMT intelligence | informa

Monetizing High-Performance, Low-Latency Networks

Low latency underpins carriers' ability to compete and win in verticals and applications

Publication Date: 20 June 2017

Ian Redpath

Introduction and research scope

Broadly defined, latency is the time interval between stimulation and response. All services delivered over a communications network are subject to latency, which is a function of distance, network architectures, network equipment, service-processing algorithms, operational procedures, and other items. Latency impacts both consumer and enterprise services.

Latency can have a direct impact on the user experience; perceptible latency has a strong effect on user satisfaction and loyalty. If a consumer notices a delay, they may opt to search for a new service provider. Demanding, technology-savvy enterprises will include latency and failover latency requirements into their requests for service. Latency performance can be the difference between winning and losing business. Latency has a direct impact on network monetization. Meeting latency requirements directly impacts latency-sensitive, high-capacity wholesale services.

Latency impacts all elements of ICT (information and communication technology), ranging from IT and cloud services, to mobile networks, consumer broadband and video services, and the MPLS network. The scope and focus of this paper is latency performance on the optical metro access, metro core, and long-haul core of the optical layer. The primary clients of this portion of the network are enterprises and internet content providers (ICPs) using high-capacity services, wholesale communications service provider (CSP) clients, and the CSPs' own retail operations business units. 5G network architectures and designs will have a tremendous impact on the access and aggregation portions of the network, but are just beyond the scope of this paper.

Monetizing low-latency networks

Inherent value of low-latency networks

Latency and its relationship to CSP revenue and network monetization is a complex topic. Latency performance underpins the entire network and therefore impacts the entire CSP business model. Potential CSP initiatives on latency could drive revenue across a spectrum of services. At the other end of the spectrum, a very tactical upgrade to improve latency on one particular service offering in a discrete region of operation could improve one revenue line item.

The transport infrastructure analogy

Economies depend on many types of transport: road, rail, sea, and air. Superior infrastructure enables national economies and catalyzes GDP. For example, factory output needs to be delivered over road and rail to shipping ports. Ports themselves need to be highly efficient and automated. Conversely, some countries struggle with infrastructure. Inferior terrestrial connections to inefficient shipping infrastructures can debilitate entire economies. In a general sense, all economies desire superior infrastructure. The challenge is in allocating investment to maximize GDP.

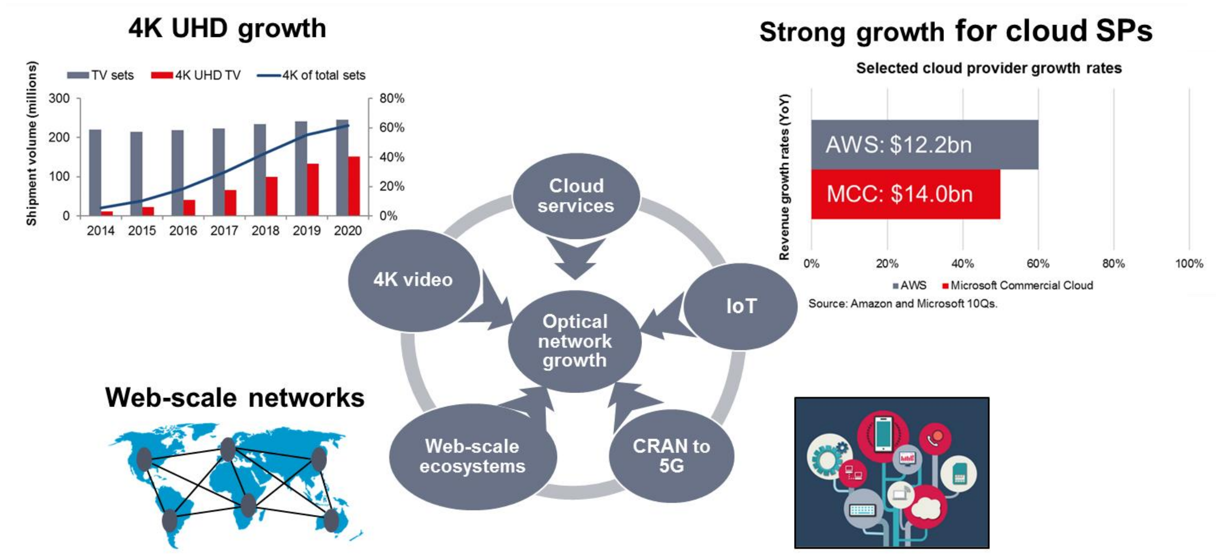
From a CSP revenue perspective, latency improvements could improve a broad spectrum of services and can potentially improve individual point services. CSPs face a multitude of challenges when determining a path for latency improvement:

- Where will the revenue of the future come from?
- What will be my optimal long-term target architecture to match future revenue streams?

- What are the near-term, more tactical business and service opportunities that would be enabled by a low-latency network?

What are the next big service and revenue drivers and where is the network going?

Figure 1: Optical network drivers and pressure to scale



Source: Ovum

Today's entire digital economy rides on the optical network and tomorrow's digital economy will place even more demands on the optical core. Briefly, next-generation 4K video will add to the network load and latency requirements. Cloud services represent a once-in-a-generation shift in IT workloads into distributed data center environments and will come with heightened latency expectations. The Internet of Things (IoT) will add billions of devices to the network with their own set of latency expectations. 5G will further add to the density of networks and will add rigorous latency specifications. Web-scale networks are global in nature and low latency by design.

Revenues are now and will continue to be generated by all the major service trends noted and by business initiatives. There is also tremendous complexity, uncertainty, and timing considerations to all the proposed business models. Drawing on the transport analogy, a low-latency network will be an imperative in the network of the future. It will be the platform for all future service growth.

Population centers and enterprise are the economic engines

Population centers are the economic engines of the global economy. The back end of the digital economy can be anywhere. The front end is where people and enterprises live. The network will need to interconnect the two in an efficient, low-latency manner. For the most stringent latency applications, network intelligence will need to reside close to the network edge. For more latency-tolerant applications, intelligence and storage can reside in more cost-efficient locations.

Networks of the future will be more data center–centric

The network of the future will be data center–centric. Data center architectures will be global, hierarchical, edge optimized, and mesh connected. Networks of the past were not designed and constructed with a data center construct in mind; this will not be the case with low-latency networks of the future. Today's networks will evolve as best they can in that direction.

Quantifying the aggregate low-latency picture

Quantifying the aggregate optimal latency network versus a suboptimal latency network would be an exhaustive exercise and would depend on an individual CSP's starting point and future business ambitions and capabilities. From a strategic viewpoint, the direction is a clear: A well-designed, low-latency network will be a strategic asset. To draw on the transport analogy a final time, a superior network, like a superior transport infrastructure, is table stakes for future successful economic growth.

Moving the discussion from the network of the future to nearer-term, enterprise-focused use cases, we find that monetization opportunities do exist that require low-latency solutions to win. A number of the use cases have been a factor in the market, will continue on, and are poised for further growth.

Enterprises have heightened latency performance expectations

Low-latency market realities

As part of the research for this white paper, Ovum conducted a channel check with leading CSPs in Europe, North America, and Asia-Pacific to gain insights into current market realities. From the CSP perspective, latency is rising in importance as a buying criterion in the high-bandwidth portion of the market. In general, enterprises and wholesale clients desire to understand latency performance on both primary and failover routes.

Rising importance of the secondary failover routes

Clients of high-bandwidth services are demanding predictable latency performance. Latency performance of the MPLS or public internet is not acceptable because routing and latency can constantly change. Clients desire traffic-engineered latency performance. In addition, clients will accept failover routes that do have a modest latency premium over the primary path but will not accept a 3x premium. For example, on a primary path between San Francisco and Los Angeles, a failover path that follows another route through California is acceptable. A backup path that routes through Denver is no longer acceptable for discerning clients.

Optical trinity: Price/bit, route diversity, and low latency

Latency is one buying criterion that has been increasing in relevance in the buying decision, but price and route diversity remain paramount. Enterprise and wholesale clients are highly sensitive to price per bit. Competition hinges on price. Clients are also much more educated on the actual primary and secondary routes. Many do want to see the physical route topology in detail and will not accept a service where primary and secondary paths are in the same physical conduit, even for short distances.

If pricing between competitors is similar, and route diversity is acceptable, latency can be a differentiator. If pricing is very different, latency will not factor in, unless there is a major latency problem. Said another way, if a longer latency is acceptable, the low price will win. Superior latency performance, not the ultra-low-latency financial case, but across general applications, can command a modest price premium. Absolute route diversity remains a key criterion in all cases.

Vertical use cases requiring low latency

Ultra-low latency for the financial services sector

The financial services sector is renowned for demanding latency performance and has been the leading vertical in pushing the boundaries of communications technology. Two fundamental elements of capital markets are market data feeds and the actual financial trading liquidity venues. Market data is at its heart pricing data for a particular financial instrument. Market data is extremely time sensitive and can vary based on source and geography. The trading venues house the matching engine that pairs buyers with sellers to execute trades. Over time, the sources of market data and the numbers of trading venues have both grown.

Additionally, in order to facilitate more transparency in financial markets and mitigate against future potential algorithmic failure, regulatory oversight is likely to increase. Financial markets have become more globalized as well.

Taken together, more market data sources, more trading venues, more regulatory oversight, and globalization fuel the need for greater connectivity for individual trading firms. Connectivity requirements have grown in concert from basic point-to-point into one-to-many. In the latter configuration, "many" has become thousands and is also global. Deterministic latency remains an imperative for primary and failover connectivity.

The global financial extranets need to tie together the trading firms with all of the key trading data centers. In order to deliver lowest-latency performance, the network needs to be constructed with trading data centers at the heart of the network design.

For the CSP community, financial markets represent a multitiered opportunity. Access remains a fundamental opportunity as does the global, deterministic, low-latency extranets or supplying regional portions of global extranets.

Financial clients are loyal to low latency but not necessarily to CSPs

The financial community remains loyal to low latency but CSPs do need to exercise rigorous and cautious ROI analysis. Taking the opportunities in turn, the global extranet opportunity necessitates a global reach. CSPs need to have building entry (fiber in the building) within the global financial ecosystem and have a network that interconnects with the entire ecosystem. Very few CSPs are actually pursuing the business at a global scale. Differentiators include buildings entered; scale of network; and ease of operations for network moves, adds, and changes.

The wider CSP community has more opportunity to play into elements within this ecosystem. The global extranet providers require local partners. Low-latency routes for portions of the global networks will be valued. In the past, CSPs have reported premium revenue for selected ultra-low-latency routes. CSPs may hold selected low-latency routes but are at risk of competitors introducing yet-lower-latency routes. A conservative approach would be to base the ROI on traditional traffic and revenue, treating the financial contribution of the premium as a value-added bonus. CSPs do hold onto these premium routes but the financial community is only loyal to low latency, not the individual CSP.

Low latency for the resource sector

The oil and gas and other resource extraction industries have an increasing need for high-performance, low-latency communications. Remote sites have basic requirements for voice and video communications to enable collaboration with colleagues in headquarters locations. Video surveillance of the operation of key assets is increasingly used in this sector, and video is used for security and safety. For example, diamond-mining operations are highly sensitive to physical and asset security; ocean-based operations are highly sensitive to personnel safety; and data-processing requirements have increased substantially in the resource extraction industry. With more and more sensors, more real-time information is collected and analyzed and utilized for control. Classic solutions for remote site connectivity were satellite based, which had bandwidth limitations and offered substandard latency performance. Today's CSP opportunity is high-bandwidth and low-latency connectivity for the emerging big data analytics and control era.

Low latency for the hypercompetitive online advertising segment

Online advertising is a growing, revenue-generating part of the market. Anyone who has browsed the internet will have been exposed to online advertising. What might be less appreciated are the underlying mechanics and the speed of events involved in the process. Internet consumers start by browsing a website or initiating a purchase such as an airline ticket. Now the consumer and their browser have been identified as a specific interest group, e.g., traveler to a certain city on a specific set of dates. As the consumer peruses additional websites, a dossier of interests builds and their profile becomes more detailed, increasing their value for targeted advertising. As the consumer moves onto the next site, further advertisement opportunities arise. Potential advertisers can bid on the ad space on that website as the page is loading. The advertiser's bid is based on the website selected and all of the known attributes assembled about the consumer. From the consumer's perspective, the new advertisement appears almost instantaneously. The websites charge the advertiser.

The back-end process involved is for each ad position; an opportunity to bid goes out to potential advertisers. Based on the user profile information, the advertisers determine an acceptable price for a specific ad and submit their bids. The winning bidder then places the ad into the ad banner as the page is loading. The enabling advertising ecosystem is based on algorithms and a low-latency network. Inferior latency performance would disadvantage bidders and reduce their business prospects. The CSP opportunity is supplying low-latency connectivity into this lucrative and growing ecosystem.

Low-latency private lines for R&E networks

"Genomics research is rapidly becoming one of the leading generators of big data for science, with the potential to equal if not surpass the data output of the high energy physics community. Like physicists, university-based life science researchers must collaborate with counterparts and access data repositories across the nation and around the globe." – Internet2 consortium.

Both physics and biomedical have the need to transmit terabit-sized datasets across intercontinental distances. Advanced techniques and low-latency networks can aid in minimizing the time to transmit datasets from computer to computer, or memory to memory, speeding up not only the transmission, but the pace of work of far-flung scientific groups.

Applications requiring low latency

Cloud service delivery – Applications need to perform as if local

Enterprises are transferring workloads to cloud service providers such as AWS, Microsoft Azure, and Google Cloud Platform. Enterprises desire LAN-like performance for the cloud service delivered over the WAN. Enterprises require a highly secure bandwidth service, a highly resilient network, and predictable, fast, low-latency performance. Early implementations of cloud access services were delivered over the public internet with typical public internet performance characteristics. An industrial-grade connectivity experience was a prerequisite to broader adoption of cloud services.

The major cloud providers operate a hierarchical network of data centers around the world. Their data center architectures include cost-optimized core data centers and distributed edge data centers. To meet the low-latency requirement, a fiber-based network from enterprise through to the edge and core data centers is ideal.

A winning and scalable cloud interconnect strategy involves a network designed for cloud access and data center interconnect. Ideally, this would include substantive fiber access into enterprise facilities, an optical aggregation network, and a data center interconnect network geared to shortest-path connectivity into key data center sites. Additionally, the network should minimize switching and routing at higher network layers and take advantage of wavelength switching to minimize latency.

High-capacity data center interconnect

The internet content provider community has been constructing a global network of data centers and their own global networks to interconnect. The most advanced models have multiple, web-scale data centers per continent with tens to hundreds of edge-cache data centers. The ICPs also desire highly robust and resilient networks, necessitating multiple links between every node on the network. At the global level, the ICPs desire multiple subsea links per ocean basin to interconnect their continental networks. The terrestrial-national-continental networks are connected with a diverse mesh. The edge-cache nodes home back into multiple web-scale nodes.

"As the world is increasingly moving toward a future based on cloud computing, Microsoft continues to invest in our cloud infrastructure to meet current and future growing global demand for our more than 200 cloud services, including Bing, Office 365, Skype, Xbox Live and the Microsoft Azure platform," said Christian Belady, General Manager, Datacenter Strategy, Planning & Development, Microsoft Corp. *"The MAREA transatlantic cable we're building with Facebook and Telxius will provide new, low-latency connectivity that will help meet the increasing demand for higher-speed capacity across the Atlantic. By building the cable along this new southern route, we will also increase the resiliency of our global network, helping ensure even greater reliability for our customers."*

The low-latency capacity opportunity for CSPs arises within the above model in all cases where the ICPs cannot build their own networks. Additionally, if CSPs can deliver a unique route, a diverse path, or a lower-latency route, opportunities will arise.

ICPs are also loyal to low latency and cost/bit

The ICP community does represent a low-latency opportunity for CSPs, but buying criteria also include price and route diversity. In terms of route diversity, ICPs require a minimum of three diverse paths out of major facilities. ICPs have become highly conversant in all of the global fiber routes. They do know the latency performance and limitations of the existing fiber routes. For some of their services and network, they are less concerned about latency, but some are highly sensitized to latency. ICPs have been addressing this issue by expanding their edge capabilities because that is where latency-sensitive services reside. Less-sensitive services and storage reside in their data centers, which have been constructed far away from population centers.

Providing unique routes that fit into ICPs' vision for network growth will be the CSP opportunity.

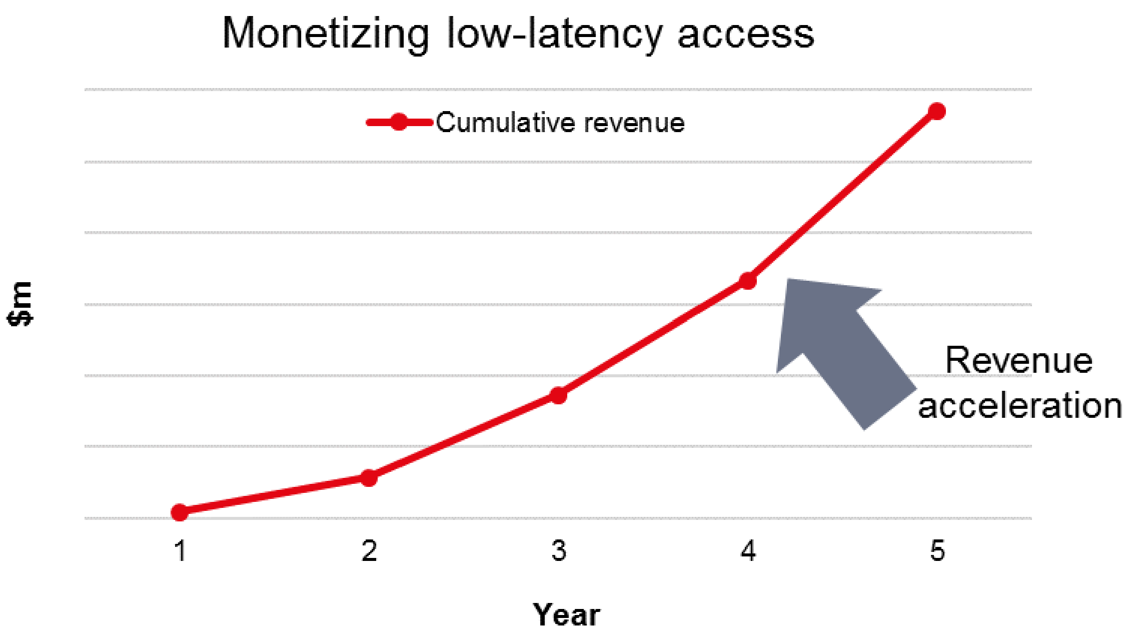
Revenue-acceleration scenarios for low-latency networks

Revenue-acceleration scenarios will range from the believable to those that require belief. The revenue-acceleration scenarios that are most defendable are the ones that are straightforward and within a defined space. As the application space broadens, it is more challenging to quantify the exact business case benefit as many more variables come into play, making it more difficult to forecast accurately.

Primary high-capacity data center access

CSPs have the opportunity to be the lead connectivity partner for every new web-scale data center constructed. The new data center can be commissioned by many entities: a large financial institution or exchange, a large enterprise, a CSP, an ICP, or a carrier-neutral data center operator. The CSP would need to be involved in the data center planning from the earliest stages. Additionally, the CSP would need a right-of-way strategy to connect the new data center and aggregation points back to other key data center properties and gravity data centers. The first CSP involved in the project with the lowest-latency network would have significant first-mover advantages. Additional CSPs entering the proposition would be in a secondary position. For a web-scale data center, the primary access opportunity could quickly become a business of multiple 100Gbps links per month. See Figure 2 for a potential revenue line item for 100G access circuits based on superior low-latency performance.

Figure 2: Primary high-capacity data center access



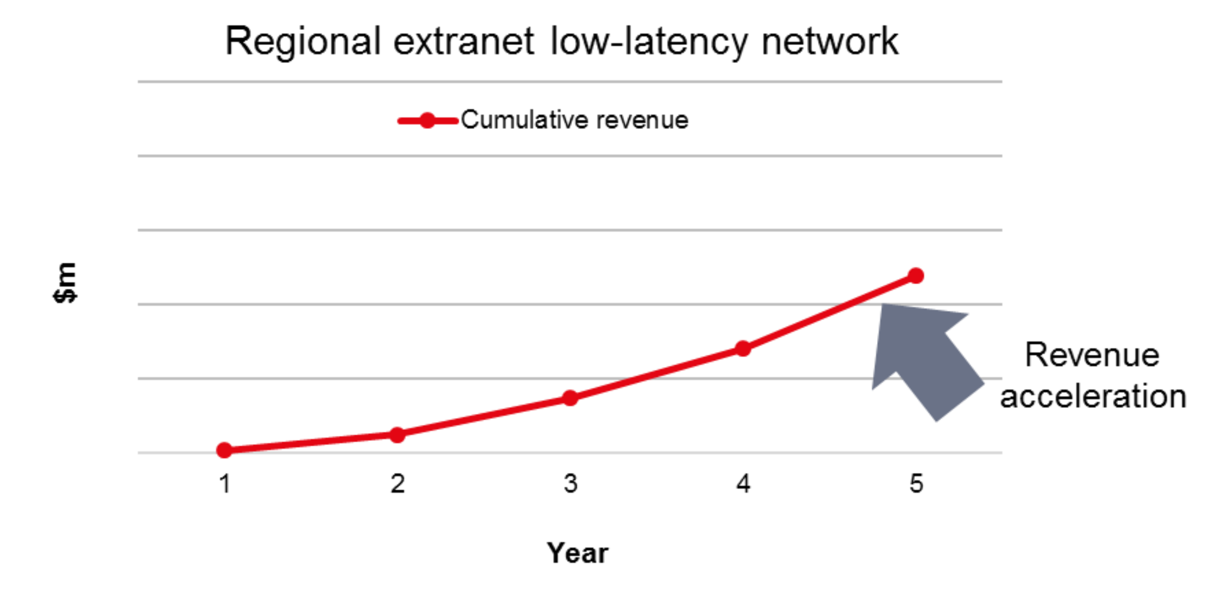
Source: Ovum

Financial extranet opportunity

The financial extranet opportunity has been an existing business for some time but continues to have an appeal for low-latency network operators. The financial community values and has a demonstrated track record of commissioning low-latency networks. Other verticals may have a general desire for low-latency networks, but may not back up the desire with a financial commitment. As noted in the above discussion, the number of nodes on financial extranets continues to grow. Additional network coverage is required.

The extranet opportunity hinges on low-latency access and interconnect. Access is required into every financial institution in the community. To achieve global access, partnerships will be required, creating the opportunity for the emergence of a global provider and regional partners. The access networks need to interconnect with a global, low-latency network. See Figure 3 for a possible financial extranet revenue opportunity based on a high volume of 1GE business.

Figure 3: Financial extranet opportunity, regional expansion case



Source: Ovum

An automated flexible network opportunity

As networks continue to grow in scale, one can envision networks of burstable packets, evolving to networks of switchable wavelengths. On the demand side, high-capacity network users will have time-based bandwidth needs. If a CSP can supply bandwidth and charge for low-latency bandwidth on a temporal basis, it may be able to create a unique market proposition.

Summarizing CSP views on monetizing low-latency networks

The CSPs involved in the study concurred that it is difficult in aggregate to quantify the total benefit of low latency from one instantiation of the network to another. However, CSPs could cite specific wins and losses based on very specific routes: wins attributed to newly constructed low-latency routes, losses due to older routes that had suboptimal market latency. The CSPs fear a deteriorating competitive position if they were not constantly striving for the lowest-latency network.

Operational practices, technology advances, and network architecture choices underpin low-latency networks

Operational practices can impact latency performance

CSPs' day-to-day operational practices can impact network latency performance. At an operational level, multiple goals and constraints have considerable influence, including deployment timelines, budgets, available equipment, and the network itself. Day-to-day practices can impact latency performance if latency is not a paramount organizational goal. Networks are built with fiber slack to facilitate future repairs in case of fiber cuts. Slack adds latency but removal of all slack would entail a greater operational effort to repair fiber cuts. CSPs do have options on slack and can emphasize its relevance to the business.

Deploying additional capacity can also impact latency. Due to the availability of network equipment and capacity utilization on production networks, capacity may be deployed on longer-latency routes if low latency is not the overriding organizational goal.

Technology advances enable lower-latency networks

A number of recent technology advances enable lower-latency network performance. Network architecture design also has a substantial impact on latency performance and a review of target architecture principles can yield performance improvements.

Coherent technology

Historically, dispersion, a fiber transmission impairment, was compensated with dispersion compensating fiber (DCF). The DCF added a 10% latency penalty. Coherent technology compensates for dispersion impairments within the digital signal processor, eliminating the need for and the latency of the DCF.

Raman amplification

Raman amplification, often used in conjunction with EDFA amplifiers, enables longer transmission distances without the need for electronic regeneration. Reducing regeneration reduces latency.

Flexible application of forward error correction (FEC) techniques

Advanced FEC techniques also enable error-free transmission performance over longer distances. FEC does add latency to transmission. For shorter distances, FEC may not be required. The flexible application of FEC techniques, matched to network demand distances, will optimize overall system latency.

Wavelength switching

Switching can be undertaken at the electrical level or the wavelength level. For a network demand between an "A" city and a "B" city, the ideal would be uninterrupted optical transmission from A to B. If an intermediate node, "C," is in the network between A and B, switching the wavelength at the C node optically would be ideal. Terminating the optical signal at C and retransmitting, switching electronically, would add cost and latency.

SDN control with integrated OTDR functionality

Networks have a greater real-time control capability than in the past. OTDRs (optical time-domain reflectometers), due to form-factor improvements, can now reside within network elements, enabling real-time measurement of network performance. Operational parameters can be sent northbound to an advanced network controller. Network policy can be set and optimized for latency performance.

Optical network architecture designs can be optimized for lower latency

The new technologies of optical switching with enhanced control, improved form factors, and low-cost optics are enabling the extension of optical network architectures in three directions:

- all-optical hierarchical architectures
- all-optical extension of the optical layer to the network edge
- enhanced wavelength switching at all of the optical nodes in the network.

Hierarchical optical networks

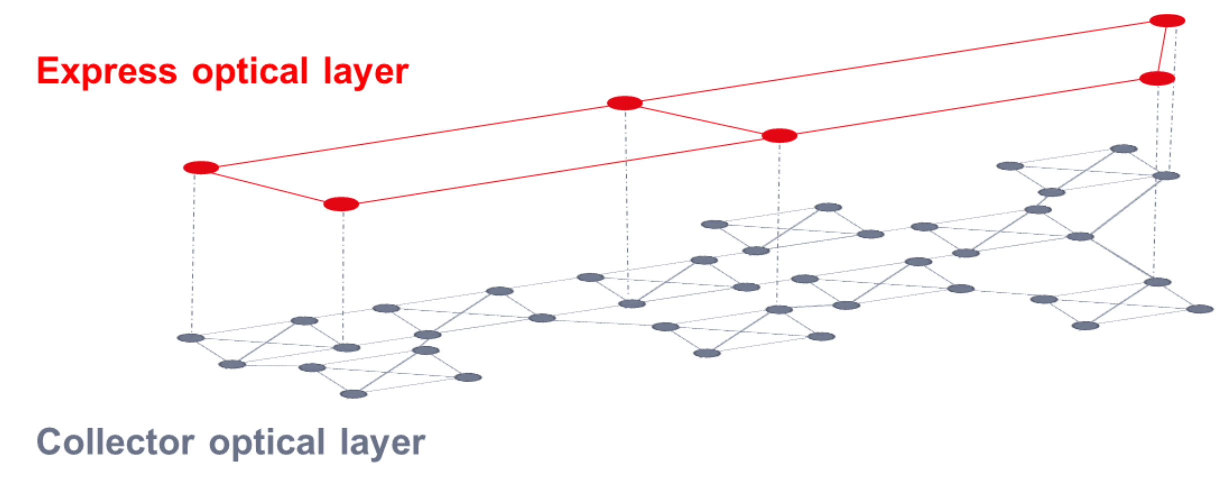
Optical networks do grow organically. Target network architectures are debated and developed. Networks are deployed and Day 1 traffic is applied to the network. So far, so good. Then things change. Traffic grows, some as predicted and forecasted, and some not. The traffic builds on the network but not necessarily uniformly. Certain high-traffic routes fill relatively quickly, while other low-traffic routes only reach a low level of network fill. CSPs can overbuild on the high-traffic routes or they can evaluate a hierarchical collector-express topology.

The preexisting network would continue in operation as the collector layer. A new express layer could be overlaid on the existing footprint. With an understanding of traffic growth from previous years and an evaluation of expected growth patterns, a provider could develop an optimal express network.

In some cases, CSPs have deployed several independent legacy collector layers at different points in time, covering the same region. Service between these layers can take weeks to provision with substantial OEO (optical-electrical-optical) conversion, resulting in large amount of potentially unstable network latency. Using the express overlay topology to connect all the legacy collector layers, a provider can create one large optical mesh network, reduce OEO and network latency, and increase the utilization of network resource.

With current-generation wavelength-switching solutions, traffic could originate on the collector network, pass through the express layer, and terminate at far-end collector nodes without terminating the optical signal at any intermediate point. The optical path latency would be shortened by avoiding unneeded OEO conversion and transiting the entire route by using a more direct express path. (See Figure 4, which illustrates the hierarchical collector-express optical construct.)

Figure 4: Express optical overlay



Source: Ovum

Extend optical networks to the network edge

Current production networks may not yet be taking full advantage of the latest optical network capabilities. Some production optical networks may stop short of the network edge, relying on more expensive cost-per-bit technologies to reach a customer endpoint. Or the optical layer may have been extended in a more piecemeal fashion over time, with multiple hops and multiple optical terminations at intermediate points en route.

Traffic from the edge has been transitioning from lower speeds to higher and higher speeds, such as 10GE. CMTS, PON, enterprise business services, and mobile networks are all evolving toward 10G-and-greater uplinks. Optical networks need to extend right to the network edge to efficiently and economically transport the traffic. The optical network edge can be a central office, customer premise, a macro cell site, or even a street cabinet. Metro optical networks may have two distinct elements: access and a metro-regional core. The access portion of the metro network can be point to point or a ring construct. 100G, 10G, and subrate services are connected to an edge optical node. The traffic mix can be aggregated to 100G wavelengths, dependent on the overall traffic demand, and then transported over the core metro network. Once on the core network, the traffic should remain optical until it reaches its destination point.

Two guiding principles of minimizing network cost are to minimize OEO conversion, which also minimizes latency, and to use the lowest network layers for grooming, aggregation, and switching. In the past, the IP layer has been used as an aggregation device but as the network scales, this becomes a more expensive proposition. Additionally, for networks with mixed traffic, e.g., Ethernet, private lines, SONET/SDH, and other protocols, OTN was designed for efficient aggregation and grooming at a lower cost per bit than a Layer-3 solution. For CSPs with OTN in the network core, extending OTN to the optical edge will be a natural progression.

Today's metro optical network elements are flexible in form factor, size, and bandwidth. Optical edge devices are scalable as well. Smaller form-factor optical edge devices can be deployed economically at the network edge without over-provisioning penalties.

CSPs can leverage optical switching, scalable form factors, and low-cost optics to extend their optical network to enterprise premises, the mobile edge, and the residential headend. The optical core can become more integrated, from the long-haul core, through to the metro network, optical aggregation networks, and multimedia access points. The integrated optical layer can support wavelength traffic end to end, optically, and minimize OEO costs and latency penalties.

Fine-tune nodal switching

Switch at the most cost-optimal and latency-optimal layer

The cost per bit for switching is the most expensive at the highest network layer, Layer 3. The cost per bit improves as switching is accomplished at lower network layers: Layers 2, 1, and 0. Latency performance also improves when switching is performed at lower layers. The selection of a switching platform is dependent on the switching granularity required. With the maturity of optical switching, wavelengths should be switched at the optical level. OTN switching can be utilized for subwavelength processing. Providers should avoid Layer-3 processing until routing is absolutely necessary.

Leverage wavelength switching

CSPs are placing a greater value on operational flexibility for multiterabit systems. Networks continue to grow in capacity and degrees per network node; automation in the core is key. Electronic switching can be used at the key network nodes but rising traffic volumes will stress electronic switching systems. Ideally, CSPs should offload electronic switching solutions for space and power savings. Wavelength-switching technology has been evolving and CSPs have more choices in wavelength-switching possibilities. Flexible grid ROADMs have achieved broadly based architectural acceptance in the market and flexible grid has been seeded into networks; many ROADM shipments are hardware-capable flexible grid. Additionally, flexible grid is required to take full advantage of all advanced modulation formats. Colorless and directionless (CD) ROADMs are also gaining broad market acceptance. For some CSPs, CD ROADMs are the right balance between additional functionality and added capex. Colorless, directionless, and contentionless (CDC) ROADMs also have been accepted by a limited number of tier-1 CSPs. CDC ROADMs do have a substantial amount of internal fiber connectivity within the network element. Manual fiber interconnection introduces operational complexity that could raise the potential of configuration errors. The optical cross-connect (OXC) has also been introduced to the market as the next stage in the evolution of optical switching. The OXC will utilize an optical backplane solution, promising a more elegant treatment of intra-nodal fiber management with greater scale.

Networks often grow organically, ultimately a blend of long-term target architecture design and practical day-to-day customer-oriented growth. The network must always be operational and up and running, but at the same time, periodically upgraded. Target architectures that were optimized for a lower-bandwidth era with earlier generations of technology can often benefit from a network refresh.

OTN switching

For the 100G era, OTN is the electrical switching construct of choice. SONET/SDH does not scale to 100G levels. For multiservice networks with a mix of packet and TDM traffic, OTN is the most cost-effective mechanism to groom and aggregate traffic for 100G transmission. As networks have scaled, OTN has scaled in tandem. OTN cluster solutions introduce a centralized switching fabric. Cluster OTN solutions can enable more efficient subrack interconnection, sparing resource sharing, saving power, and enabling more flexible capacity expansion.

CSPs need to manage the transition from electronic switching to wavelength switching. Subwavelength traffic can be processed via an OTN switching platform while wavelengths can be switched via an optical platform.

Conclusions and recommendations for CSPs

Optimizing the network for latency

Optimizing the network for low latency is a wide-ranging topic. Solutions and recommendations ultimately will vary widely by CSP, dependent on ultimate business goals, target network architecture, network starting point, capital budget available, and other unique situational characteristics. Recalling the technology discussion above, solutions can range from a modest software change to a complete network over-build. Solution timescales can range from minutes to decades. Capital requirements will also vary widely. How does one begin? What investments can be made that will yield return and not strand capital? What are the ongoing industry best practices? There is a list of potential actions that can be undertaken to optimize the network for low latency including:

- Operational practices level
- Product level
- Network architecture level
  - Maximize optical bypass
  - Extend optical transport network to the network edge
  - Use wavelength switching
- Fiber routing level
- Network scale level

All the possible latency-improvement possibilities are business objective and ROI dependent and will have differing timelines based on how easily they can be introduced into the network.

Review operational practices and procedures

Refinements for day-to-day operational procedures can yield latency improvements. Fiber slack policies can be reevaluated and applied with more rigor as the latency goal becomes an organization imperative. Capacity deployment procedures can be reviewed and reevaluated with low latency as a key operational goal.

Product level

At a minimum, most CSPs are vision aligned with coherent networks. Coherent networks continue to be rolled out at scale, eliminating the latency from DCF. CSPs have also been deploying vendor solutions that are optimized for latency. In the near future, CSPs will have more dynamic control over FEC as well. For these types of latency improvement options, CSPs are fairly quick to adopt and implement these solutions into the network.

When revamping access networks, update the access and aggregation portion of the optical network

The next set of recommendations are more challenging and costlier to implement, but if enacted will improve latency performance. An alternative way to improve latency is to eliminate OEO conversion. All CSPs desire to remove OEO from the network because they immediately associate OEO optimization with capex savings: remove unneeded OEO, remove cost. A byproduct of optimizing the network by minimizing OEO is that latency performance will be improved.

When a major upgrade to the access network occurs, CSPs can consider updating the access and aggregation portions of the optical network and extend the optical network to the edge. Again, very easy to say, but capital- and time-intensive to enact. Updates to the access, aggregation portion of the optical network may be conducted in concert and even necessitated by access technology upgrades. One example would be how 4G deployment drove an optical upgrade cycle. Similarly, 5G is likely to also drive a major upgrade cycle in the optical edge and core. Capital budget allocation for optical network upgrades on their own could potentially be challenging. Optical upgrades paired with revenue-generating access technology upgrades has been a path forward in the past for many CSPs.

Commission a new optical layer including new optical lines with optical switching nodes

Commissioning new optical lines will be a possibility for some CSPs on multiyear timescales. CSPs will fill their existing operational networks, but as those networks fill, planning for new network overlays will be a possibility. When the time comes to plan and build a new line, low latency will be an important network design parameter.

Recommendations will ultimately vary by CSP, dependent on business models, target architectures, starting point architectures, available capital, and other considerations. For individual CSPs, the exact timing of the build evaluation and implementation will play a key role as well. Technology will continue to advance. There may be a promising technology that will be available in 4–5 years, but networks have to be built now. CSPs build in cycles and technology advances in cycles. Optics has recently had a number of point advances that have added up to a new era of optical networks. 100G transmission, electronic dispersion compensation, programmable FEC, and Raman amplification have been in production networks for some years now. Flexible grid networks, advanced modulation formats for beyond-100G transmission, and industry-proven wavelength switching have recently been validated and deployed in production networks with advanced software control and are now ready for wider-scale deployment.

Wavelength switching has smaller form factors and more elegant fiber management and is now ready for wide-scale deployment.

The new-generation optical layer is inherently lower latency, due to a number of advances:

- Latency-lengthening dispersion-compensating fiber is gone.
- Reaches have been lengthened and with wavelength switching, OEO conversions have been eliminated and latency reduced.
- FEC is configurable; for shorter network demand cases, FEC is not needed and shutting off FEC will reduce latency.

CSPs with refreshed networks will have a competitive latency advantage over those operating older-generation networks.

OTN switching's role in the network has increased

With the widespread adoption of 100G in the network, OTN switching has become the "go to" mechanism for multiservice network grooming, aggregation, and switching. CSPs can consider "right-sized" OTN grooming, aggregation, and switching for multiple situations in the network. At the network edge, OTN can be utilized for aggregating and grooming traffic onto 100G wavelengths. In the core switching nodes, OTN can be used in the switching and re-grooming application. If scale is required, cluster OTN can be deployed.

Optimize the fiber path

Beyond short-reach metro distances, the fiber route immediately becomes a substantial contributor to latency. Many CSPs – quietly without a lot of fanfare in industry press – extend their fiber networks year after year. Some leading CSPs report fiber-fed enterprise buildings entered. Some report fiber miles. Leading CSPs continue to add to the stock of the fiber footprint.

The fiber-extension efforts will vary widely in business value, cost, and time to build. Beginning at the metro end:

- Short-reach fiber laterals to new properties off of existing fiber networks.
- Metro extension builds: CSPs will maximize value by targeting data centers, enterprises, macro cell towers, and PON headends.
- Regional builds to add underserved territories and potentially add remote data center properties.
- National builds in developing economies or a new "rail" in developed economies.
- Subsea network links to create more direct and low-latency paths between key sites across continents.

CSPs have also undertaken specific initiatives to lower latency for key network nodes via fiber overbuild, including the following approaches:

- Classic fiber routing, which is city center to city center. New data centers have been commissioned in suburban regions. Following the city center–to–city center path and then out to the suburbs creates a longer "trombone" path. CSPs have overbuilt fiber routes following a more direct path to remove the trombone and optimize the latency.
- On long-haul routes, city center to city center followed the historical railway routes, adding latency. Between major financial centers, the secondary cities added latency but not commercial value. CSPs overbuilt, bypassing smaller cities and lowering latency.
- CSPs have built network bisectors to reduce latency.

The ICPs have commissioned multiple transatlantic and transpacific cables to lower latency between their major data centers of operation.

Looking ahead, more data centers will be commissioned. There will be a set of leading high-value data centers that will be high-value nodes on the digital economy landscape. CSPs need to track the ongoing development of high-value sites and plan their network expansion accordingly.

In the major data center cities of the world, new CSP entrants have commissioned brand new fiber routes, not following the classic CSP network footprint, but charting a network footprint optimized for low-latency data center interconnect. The new CSPs identified all the high-value assets (new suburban data centers, leading enterprises, and gravity data center interconnect nodes) and built their network to serve the low-latency demand. Some nontraditional rights-of-way tactics were included such as aerial fiber and fiber routed through waterworks.

CSPs, as with every business, are resource constrained. Network-planning tools have evolved. With latency becoming a more important network design parameter, CSPs can now take advantage of the latest capabilities in network-planning tools. CSPs can leverage greater intelligence and capabilities within network-planning tools to design and deploy a more latency-optimized network. Via the planning tools, CSPs can run multiple scenarios in fiber build and route deployment cases, to determine the lowest-latency configurations.

The above set of latency-reducing options does vary in cost from hundreds of thousands to hundreds of millions. Individual CSPs must track and forecast the growth of high-value network nodes and plan network extensions to meet the opportunities.

Scale the network

Many CSPs have been following this long-term strategy by both organic and nonorganic means. Most CSPs continue to steadily add to their network footprint by adding fiber nodes and additional links. A number of CSPs expand through ongoing acquisition, continuously buying other CSPs to gain an ever larger footprint. More fiber-entered buildings with a denser mesh network enable more paths through the network, creating more latency-optimized options.

Appendix

Author

Ian Redpath, Practice Leader, Components, Transport and Routing

ian.redpath@ovum.com

Ovum Consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Ovum's consulting team may be able to help you. For more information about Ovum's consulting capabilities, please contact us directly at consulting@ovum.com.

Copyright notice and disclaimer

The contents of this product are protected by international copyright laws, database rights and other intellectual property rights. The owner of these rights is Informa Telecoms and Media Limited, our affiliates or other third party licensors. All product and company names and logos contained within or appearing on this product are the trademarks, service marks or trading names of their respective owners, including Informa Telecoms and Media Limited. This product may not be copied, reproduced, Whilst reasonable efforts have been made to ensure that the information and content of this product was correct as at the date of first publication, neither Informa Telecoms and Media Limited nor any person engaged or employed by Informa Telecoms and Media Limited accepts any liability for any errors, omissions or other inaccuracies. Readers should independently verify any facts and figures as no liability can be accepted in this regard – readers assume full responsibility and risk accordingly for their use of such information and content.

Any views and/or opinions expressed in this product by individual authors or contributors are their personal views and/or opinions and do not necessarily reflect the views and/or opinions of Informa Telecoms and Media Limited.

Ovum

TMT intelligence | informa

CONTACT US

www.ovum.com

analystsupport@ovum.com

INTERNATIONAL OFFICES

Beijing

Dubai

Hong Kong

Hyderabad

Johannesburg

London

Melbourne

New York

San Francisco

Sao Paulo

Tokyo