Tutorial on the Design and Construction of Local and Regional Exchange Facilities

Version 0.3
March, 2001
Bill Woodcock
Packet Clearing House

This tutorial addresses the questions and issues associated with the formation of local and regional Internet traffic exchange facilities. When and where are they needed? What are their physical and infrastructural requirements? What business model is most appropriate, and how can you finance the costs? What services should an exchange point provide to its users, and what policies can be established to ward off trouble?

Determining Need

While it's easy to assume that enthusiasm and resources are all that are needed to get a new Internet exchange off the ground, it's imperative to first turn an objectively assess whether a need actually exists. An optimum distribution of peering points relative to the size and density of user population and telecommunications price/distance sensitivity obtains, and creating additional peering points beyond that optimum density increases market confusion, raises costs, and decreases value for everyone involved. Conversely, a too-sparse distribution of peering points also increases costs and decreases user-perceived performance and value.

First, it must be determined whether a sufficient end-user population exists within the locality of the proposed exchange point to support a local exchange of traffic. Since the primary function of exchange points is to shorten delivery paths between nearby sources and destinations, if an insufficient number of local parties are sending traffic *to each other*, rather than to remote destinations on the Internet, an exchange point won't be viable.

Furthermore, if there's a preexisting exchange point in the area, it's never economically beneficial to create a new unassociated one, rather than either fixing any problems that may exist with the current one, adding facilities to it, or annexing it to a new but fully interconnected facility. Political, philosophic, economic, or technological differences with the management of existing exchange facilities often seem like sufficient motivators for the creation of a new exchange, but in fact these are the very worst of reasons, and any new exchange created in this way is doomed to fracture the market from its inception.

Geographic Location

If you determine that a new exchange facility is warranted, site selection is of crucial importance in making the exchange a success. Data communications infrastructure exists in an odd state of compromise, partaking partly of the nineteenth-century world of rights-of-way and governmental authority, partly of the twentieth-century world of multinational corporate ventures, and partly of the twenty-first century world of virtuality and cyberspace. The competing demands of each of these collide when siting exchange facilities.

A community of users must exist, served by multiple Internet providers, economically separated from the nearest existing exchange by high telecommunications backhaul costs, and as stated before, communicating amongst themselves with sufficient frequency to make the

overhead costs of local interconnection preferable to the scaled costs of backhaul.

A nexus of fiber, copper, or rights-of-way must exist within or proximal to the user community, and it must be open and accessible to those companies which wish to add to or improve upon those facilities.

Lastly, a physical space, a building, portion of a building, or directly adjoining cluster of such, must exist at that nexus, and be accessible to companies which wish to use the fiber, copper, or rights-of-way to make Internet services available to the local user population.

When all of these conditions exist, a frequent result is a “telco hotel” or a building which has been stripped of human amenities in order to render it more suitable for telecommunications termination and interconnection. Telco hotels and large governmental and institutional telecommunications consumers are by far the most frequent hosts to Internet exchanges.

Telecom hotels may advertise themselves and be easily found, or they may be concealed behind unobtrusive facades in unlikely-seeming areas. The surest way to find them is to find the intersection of telecommunications rights-of-way and then look at the tenant-lists of adjacent buildings. Rights-of-way documentation and “as-builts” (after-the-fact engineering drawings of installed facilities) are often available as public records through municipal planning departments, and some pathway maps may be available to customers or potential customers of telecom providers through the providers’ provisioning departments. Lastly, in areas where new fiber is being installed for the first time, opportunity may exist to artificially influence the creation of an appropriate interconnection facility by working with governmental bureaus of economic development and municipal planning to require new pathways to cross through one or preferably redundant spaces where appropriate real-estate is available.

Density

Although exchanges which exist in a single contiguous space within a single building are by far the most common, several alternatives and outgrowths of this model exist. A common development in exchanges which prove successful over time is the extension of the exchange facility to a cluster of adjacent or nearby buildings, each under different business management with differentiation in services and pricing, but all sharing a common switch fabric which allows all their customers to interconnect.

Less common, because they provide less economic leverage, are distributed exchanges which utilize a “MAN” or metropolitan area network, or a frame relay or ATM switch “cloud.” Such models are dependent upon ubiquitous availability of telecommunications facilities at a low price, since they require that each participant have the ability to interconnect with all others from their own preexisting location. The first commercial exchange, the MAE or Metropolitan Area Ethernet, built in 1991 and 1992, interconnected four providers’ facilities in different parts of Washington D.C. with FOIRL, ten-megabit Ethernet over fiber. Likewise, the first exchange built by Packet Clearing House in 1993 and 1994 was based upon frame-relay, utilizing spare capacity within a dozen providers’ preexisting interconnections to the PTT’s frame-relay switch fabric to create a distributed exchange within the Northern California region.

So-called “layer 3 exchanges” also exist, in which each provider peers with a central route-server, and thus transitively with each other, but have no direct layer-2 connectivity to each other’s routers. Route servers are beneficial when used in conjunction with a layer-2 exchange, but have historically been subject to much abuse when attempts have been made to use them to the exclusion of layer-2 connectivity.

Hybrid infrastructures also exist, and can often provide useful flexibility. When MFS built a second MAE facility in San Jose, it was joined to the Packet Clearing House frame-relay exchange point with an Ethernet interconnect which allowed MAE participants to peer with PCH participants and vice-versa, and allowed participants the liquidity to choose or move between the two business and pricing structures without forfeiting their peering connections. A similar conjoining exists between the PCH frame-relay exchange in San Diego California and

the San Diego Network Access Point, or SDNAP, an Ethernet and FDDI exchange at the San Diego Supercomputer Center. A different combination, the CIX, or Commercial Internet eXchange, a layer-3 exchange, and the PAIX, or Palo Alto Internet eXchange, each became more successful after collocating and interconnecting with each other.

In general, though, the all-participants-in-one-room model is most common because it's simplest and usually provides the greatest value for its cost.

Building Management

In the case of a physically centralized exchange, the ownership and management of the building or buildings in which it's housed can have a substantial effect upon the success or failure of the exchange. Obviously benevolent and well-informed management is of great benefit. Exchanges have been hampered or forced to move or close due to antagonistic relationships with building management which was not inclined to make the considerable allowances necessary for telecommunications, power, and cooling facilities penetration into a building, or which didn't understand the business model and needs of an exchange sufficiently to protect it from inadvertent encroachment. Exchanges have also been ruined by too much attention from building management which perceived them as a lucrative revenue source.

Telco hotels are often owned by real-estate investors, and managed for maximum revenue extraction. The prices borne by users of densely-populated telco hotels are extraordinarily high by comparison with those which retail or office use of the same building would produce, often by a multiple of eighty or more. The cost of maintaining and upgrading power, cooling, and riser facilities, however, is not insignificant either. Professionally managed telco hotels and colocation facilities can be among the most convenient spaces to use, if the cost is justified by a plentiful exchange of traffic.

Universities and public research institutes, as large noncommercial and usually nonpartisan consumers of telecommunications services, also often host exchange facilities. Typically a space set aside within a university's own datacenter or telecom entrance facility, these are often moderately well furnished with cooling, power, and fiber providers, yet relatively inexpensive since unlike a real-estate investor, a university can derive benefit directly from the hosting of an exchange, in that it can participate in the exchange as a peer, with a zero-mile backhaul cost.

Some municipal governments are cognizant of the same benefit and have subsidized the formation of exchange facilities as a means both of enhancing their own connectivity internally and to their citizenship, and of spurring the growth of Internet, high-tech, and communications business and consequently tax-base within their borders. Such facilities are often sited within a municipality's emergency-services datacenter, that being the closest thing cities typically have to appropriate facilities. Boston and Kyoto host notable and successful examples of this type of Internet exchange. Government-owned exchanges are typically both helped and hindered by their association, in that bureaucracy surrounding rights-of-way and antenna-siting are often waived or made exception to, but other policy dictates may intercede in the operation of free market forces within the exchange.

Facilities

Centrally constituted exchanges typically have four major infrastructural requirements: pathways, power, cooling, and security. Each may have some or no measure of redundancy, and each may be plentiful or constrained. Without all four in approximate balance, an exchange is unlikely to be able to provide value to its users.

The word "pathways" refers to the space required to run fiber and copper from outside building penetrations through the building to patch-bays and equipment which requires connectivity. In multi-story buildings, this consists of horizontal conduit, typically run against the ceiling of each level, and risers, or vertically aligned penetrations through each floor,

through which cable can be run up and down between levels. Capacity, redundancy, reuse, and firestopping are the major concerns in maintaining pathways.

Making pathways sufficiently large in the first place is surprisingly difficult, both because of structural and firestopping requirements in the building's engineering, and because building-owners nearly universally underestimate the portion of their building which should be dedicated to pathways, which are typically considered overhead rather than directly revenue-producing.

Two risers are often provided at opposite sides of the building, isolated to the greatest degree possible both in terms of likelihood of simultaneous failure in the event of structural collapse of the building, and in terms of the likelihood of fire spreading through both simultaneously.

Reuse is a constant concern in successful facilities, where conduit and riser space are always at a premium. The most effective method of reusing pathways tends to be high-density in-building cable-plant owned and maintained by the building management, and terminating on patch bays as near to the entrance facilities and to the end-users as possible. Thus cable need not be removed or re-pulled through conduit and risers, for the most part. Failing that, dedicated cables, which typically cannot be reused, need to be marked at least at both ends, and preferably periodically along their length, and tracked in a database such that they're associated with the parties at each end. If either of the parties leaves the facility or requests a disconnect, the entire length of the cable must be removed from the pathway, thereby allowing the reuse of that portion of the cross-sectional area of the pathway. It's also generally felt that some recurring housekeeping fee should be assessed for each cable in the building pathways, to provide users with an incentive to request disconnects when cables are no longer truly in use.

Firestopping is the science of preventing fire from spreading through a structure. This is typically done by using materials like concrete, which tend to maintain their structural integrity during a fire and prevent the passage of flame, and by limiting the number, size, and alignment of penetrations through elements of the structure. This need is in direct opposition to the need for easily reused cable pathways through the building. The compromises which are typically arrived at are the use of huge barn-like single-room structures, the partitioning-off of risers such that flame which travels within them does not have direct access to each level which they pass through, and most commonly the resealing of penetrations using reenterable fire-retardant wadding, putty, or sealant. Two related areas are fire suppression and flood control. Many types of fire-suppression systems are available. Those which douse critical electrically-powered equipment with water are frowned upon. Since exchange facilities are generally unpopulated, chemical fire-suppression, while more expensive, is preferred. If a water system must be used, a "dry standpipe" system, in which the pipes are filled with pressurized gas rather than water under non-emergency conditions, is preferred, since it cannot leak water onto equipment below. In many facilities, water-based fire-suppression systems are illegally capped-off subsequent to construction inspections, since it's usually economically preferable to allow equipment to operate until it's destroyed by fire, rather than destroy it earlier with water. Whether water comes from a fire-suppression system or from without, it's important that it be given a safe route out of the facility which does not put further equipment in its path and thus in danger. Flooding is one of many kinds of natural disaster which may affect the choice of overall site, as well.

Power and cooling are also critical to the success of an exchange. Computers and telecommunications equipment have become the single largest consumer of electrical power in many developed countries, and a good part of that demand is concentrated in exchange and colocation buildings. Facilities which were built a few years ago to provide fifteen amps per rack cannot be filled, because they ran out of power long before they ran out of rack space. Newer ones, with two twenty-amp circuits per rack are starting to see the same problems as power-hungry PCs and RAID arrays are reduced to 1U in size and are being installed forty to a rack. Power must also be made reliable through battery strings and backup generators or cogeneration. High-end facilities tend to have battery strings which can power the entire facility for perhaps fifteen to twenty minutes, a pair of redundant backup generators, each capable of powering the entire facility, and two or more independent utility power-grid connections. Ideally these constitute two entirely separate systems, such that users of the space can request "A" and "B" circuits into their rack, and be assured that there aren't significant common points of failure between the two. More power-hungry facilities operate cogeneration plants, producing power continuously. This has the benefits of independence from the vagaries of grid power, lower per-watt/hour cost, and importantly, a heat or steam source which can be converted to cooling.

The cooling of thousands of machines is a not insubstantial engineering problem in itself,

particularly when the machines are air-cooled and there's no uniformity to the direction of airflow through them. Some calculations show that the airflow through a closed rack of 1U servers now must be maintained at nearly 60 miles per hour. Network interconnection equipment isn't quite as hungry or hot, but it's headed in that direction, and more of the equipment at exchange points belongs to "content peers" with servers rather than routers, as time goes on. Again, a side-benefit of cogeneration is that the exhaust from the turbines can be used to drive thermal-transfer chillers. While in a backup-generation situation the generators must be sized at approximately 1.75 times the electrical load of the communications equipment, in order to accommodate the air-conditioning for the equipment, cogen needs to be only one times the electrical load, since the air-conditioning is a byproduct of its normal operation. Siting an exchange facility within a building which already provides some or all of these amenities is a huge work-saver.

Security is probably the area in which the widest variation is tolerated. At a minimum, a locked closet in a building which has the outside doors locked at night, and one set of keys issued to each participant, or keys kept by a tenant or manager of the building, may be sufficient to keep a small exchange point out of trouble. At the opposite extreme, there are facilities with multiple biometric authentication devices and continuous escorts within the facility. Many exchanges compromise with a card-key system that logs who's entered the facility when, and perhaps restricts access to individual locked cages or enclosed racks. The question of whether to allow 24-hour access to all participants is often a defining one for small exchanges, since it's a requirement for many larger potential participants.

Services

The primary service of any exchange is to facilitate the interconnection of the participants. This is generally done primarily through a common switch-fabric, supplemented by individual cable crossconnections between specific pairs of high-traffic peers.

There have been experiments with a good many types of peering networks, starting with shared ten-megabit Ethernet, all the peers in a common collision domain, which quickly gave way to switches, first 10Base-T, then FDDI, then 100Base-T, and now Gigabit Ethernet. Soon large exchanges will be using 10GigE. Historically, there have also been attempts to use ATM as a switch-fabric medium within centralized exchanges. The management benefits have been touted, but ATM has proven too fragile, too inflexible, and too expensive for mainstream exchanges. It's also generally limited to 155 megabits or 622 megabits at the very fastest router interfaces, and there's little pressure on equipment vendors to do future development on ATM, so it will probably fade away more quickly than FDDI did.

Private crossconnections between peers are typically performed on 100BaseT copper, or GigE fiber. As mentioned in the discussion of pathways, well-run larger facilities, these crossconnects are labeled at both ends and tracked in a cable-management database, so that they can be removed as soon as they're no longer in use.

The labor of installing crossconnections, as well as physical tasks within the facility, are often accomplished with "remote hands" service. That is, an on-site technician in the employ of the exchange, who is available to perform tasks within the facility on behalf of its participants. Most small facilities do not have formalized remote-hands services, usually relying upon members who are nearby or work in the same building, whereas most larger ones have technicians available for this purpose around the clock.

Most exchanges have a route-server or looking-glass, that is, a BGP-speaking device which peers with all of the participants and collects route advertisements from them, either for the purpose of reflecting those routes to the other peers (a route-server) or for debugging, diagnostic, and research purposes (a looking-glass).

Some exchanges provide a variety of ancillary services beyond these staples. Most common are the maintenance of a GPS-synchronized or atomic clock to act as a stratum-1 NTP time server and a web-cache which participants can query for web pages before trying their upstream providers.

Business Structure

The legal form in which an exchange is embodied can have a great effect, positive or negative, upon the exchange's long-term health and success. One of the first choices to be made is whether to form a corporate entity to own and operate the exchange, or leave it an unincorporated entity. In most countries, incorporation resolves a lot of issues regarding ownership of the physical property of the exchange, any contracts which it needs to enter into, individual volunteers' liability, and the ability to employ staff. On the other hand, incorporation generally commits the organization to significant ongoing overhead costs in both money and bureaucratic procedure.

If the exchange is incorporated, this raises the issue of ownership, which is key to setting and maintaining participant, community, and regulatory confidence in the durability and neutrality of the exchange. In short, exchanges are most successful when all parties are ensured that control of the exchange will not fall primarily into the hands of any one interested party, or into the hands of a cartel. This tends to argue in favor of one of three ownership models: cooperative, external, or self-ownership.

In a cooperative, each participant in the exchange is issued a share and a vote, both nontransferable and revocable upon their withdrawal from the exchange. The more successful the exchange, the more broadly ownership will be spread, and so the less likely the exchange is to fall under the control of any one party. On the other hand, if a cooperative fails to delegate decision-making tasks from the full voting body to staff, it can become a quagmire, preventing the exchange from reacting with necessary alacrity to external changes of market, technology, or demand, and ultimately stunting the exchange's growth.

Under external ownership, exchange facilities are owned and operated by a third party which is neither a participant in the exchange, nor a competitor, nor otherwise supplier to any of the participants, nor likely to be acquired by such. For example, universities, governments, and real-estate investment banks are all common exchange-point operators. The primary advantage of external ownership is that it relieves the membership of organizational governance responsibilities and may provide some initial external investment. The disadvantage is that it may tie the future fate of the exchange to a third-party's profit motive, thereby putting the exchange at risk of coming under the control of an interested party, or of excessive capital extraction.

Self-ownership means incorporation as a non-profit entity. With very few exceptions, non-profits are corporations which own themselves, rather than being owned by shareholders. Non-profits generally receive legal protection against acquisition by anything but other non-profits with similar charters. This fully relieves the risk of concentration of ownership, but requires that the membership form a governance structure similar to that of a cooperative, so that individual members don't have undue sway in the selection of staff and the setting of policy.

The final option, an exchange which is owned by one or a small cartel of its participants, or which allows the transfer of ownership shares so that such a concentration of ownership could occur, is to be avoided at all costs. The specific problem is that in an exchange of, for instance, ten participants, in which each participant must make an investment of time and resources to participate and build the exchange, even if all begin as equals each participant has only a one-in-ten chance of becoming the final sole owner of the exchange. Thus all potential participants are disincented from making the necessary initial investment, because the most likely outcome of doing so is that their investment will be converted to the benefit of one of their competitor/co-participants. Simultaneously, while the potential advantage gained by the one participant who manages to accrue a controlling interest is greater than the investment they've put into it, it may not be ten times that amount, so participants aren't even greatly incented to participate on the chance of becoming the eventual owner of the exchange. In either case, fungible ownership of an exchange creates an atmosphere of low confidence, in which participants and potential participants are generally unwilling to make the investments necessary to the eventual success of the exchange.

A common course of development in regional and local exchanges is for a group of potential participants to self-organize and determine a need. Next, prior to forming a legal entity, they informally divide the necessary work and expense, and create the exchange in its initial form. This is typically done with a minimum of investment, so it rarely has a lot of amenities or

redundancy or features above the bare necessity of a mutually-accessible layer-2 switch fabric. It's typically left to operate in this mode until upgrades are necessitated by increased demand for features or throughput or port capacity, whereupon it's incorporated as a non-profit or a jointly-owned cooperative or a combination of the two, to provide participants with the reassurance they need prior to making a second and usually more substantial round of investment. Thus the initial unincorporated phase serves as a low-cost "proof of concept" and the increased overhead of governance of an association is delayed until the value of the exchange has justified it.

At some point during the growth of the exchange, it should become obvious whether staffing is necessary. If the exchange is externally-owned, the parent may have administrative resources which can accomplish tasks like billing and purchasing, and technical work may be more easily performed by the members. Many cooperative and non-profit exchanges outsource administrative "front office" work, and operate without any dedicated staff. Larger ones, however, will generally need to employ both administrative and technical staff.

Lastly, the cost of participation in the exchange must be determined. If the form of the exchange is either a non-profit or a cooperative, the price of membership and participation is most likely to be determined as a function of cost-recovery plus some margin of safety. This can be calculated either predictively, or after-the-fact based upon actual costs once an operating buffer has been established. If an exchange is unincorporated, its costs are usually informally covered by whichever participant seems most motivated to do so at the time they're incurred. If the exchange is operated by a third-party, it may charge on a cost-recovery model, if that third party has reason for benevolence, or it may attempt to set or follow market pricing.

Policies

The success of an exchange point is also greatly affected by the operational policies by which it governs members' participation.

A number of technical policies for switch-fabric interconnection are widely agreed upon at this point. The first of these was "no pointing default." This means that participants agree to only send solicited traffic to their peers across the exchange. No participant may send data traffic to another participant if the destination address is not one they've explicitly learned via a peering session, directly or through a commonly-agreed-upon route-server, from that specific peer. This prevents participants from using static routes or route-maps to direct traffic to other participants who haven't agreed to peer with them. It also prevents a specific method of theft of service whereby party A, which peers with party B at multiple locations, points static routes at party B's routers, enabling the formation of tunnels through party B's backbone between the exchange points. This attack has been historically common enough that peering point participants should protect themselves against it explicitly, however, rather than relying upon policy compliance by their peers.

A common prohibition against directed broadcasts requires that participants not propagate packets from the outside which are addressed to the exchange point switch fabric's subnet broadcast address, thereby preventing many classes of denial-of-service attack from being propagated through the exchange.

A prohibition against proxy-arp and ICMP redirects across the exchange generally limit the damage which misbehaving routers can inflict upon other participants' operation. Proxy-arp would allow a router to request that traffic which it hadn't solicited via a BGP advertisement, while ICMP redirect would allow a router to redirect a route's next-hop to a third party which hadn't solicited it.

A common prohibition against the use of auto-detection of speed and duplex-settings on equipment interfaces facing the switch fabric is the result of lessons learned from the frequent failure of auto-detection to make correct choices. The speed of an interface should always be statically configured, and should always be full-duplex.

There are also occasionally prohibitions against the transmission of non-IP protocols like IPX,

AppleTalk, and DECnet, against the use of IGPs like OSPF and IS-IS, or against the use of discovery protocols like CDP. While these don't have specific problematic failure modes associated with them, limiting them is considered good-housekeeping by some.

A few other technical policies, primarily related to extensibility of the switch fabric, are subject to disagreement, and different exchange points have chosen to implement them in different ways. The central point of contention is whether participants or other exchanges should be able to extend an exchange point's switch fabric by adding switches to it, outside of the administrative domain of the first exchange point itself. Exchange point operators commonly argue against doing so, either from a position of ignorance of the economic rules which govern their business and a misunderstanding of the nature of competition in the exchange point marketplace, or quite legitimately on the basis that it decreases the robustness and reliability of the switch fabric. Specifically, some exchange points ban, as a group, the use of more than one MAC address per switch port on the exchange, the transmission of spanning-tree packets, the connection of anything other than a router to a switch port, and the use of more than one switch fabric IP address per switch port. Following these rules makes for a very simple exchange, which is certainly unlikely to fall prey to layer-2 instability, but it also creates a crippling set of artificial barriers to entry, liquidity, and competition, decreasing general confidence and customer trust by jeopardizing free operation of the marketplace. Furthermore, it lowers the value available to customers within the exchange, by partitioning them from peering with participants who are within nearby facilities. This topic is too deep to fully engage within the scope of this tutorial, and will be the subject of a separate white-paper.

Another set of interrelated policies surround the issue of data collection, ownership, and dissemination. Many exchange points require each participant to peer with a common looking-glass device which contains information accessible only by the exchange point operator, or only by the exchange point's participants. Most also have a non-mandatory looking-glass which is visible from the outside. Some require full routes, some require customer routes, some require both with communities set to differentiate them. Some exchange points make switch-port statistics available only to the user of a switch port, while some make them available to all participants. Some make aggregate information available to participants only, while many make aggregate information public. When exchange points suffer technical failures, some disclose that information fully, while others disclose it only to participants, with a requirement that it be kept in secrecy. The more open an exchange is with its intellectual property, the more transparent the market, and the greater the eventual value of the exchange will be, but it must weigh that against the possibility of discouraging naïve participants who believe that they can gain some benefit from secrecy.

Lastly, a critical decision which faced exchange points historically, whether to support bilateral peering only, have a multilateral peering agreement, or a required "mandatory multilateral peering agreement" has been largely worked out by this point. Non-compulsory multilateral peering agreements are a great convenience to members, and there's no reason not to have one. Mandatory multilaterals attempt to compel user behavior to a degree that has not been supported in the marketplace, and new exchanges which have attempted to require them have failed in the face of open competition.

Acknowledgments & Attributions:

This paper draws upon documents provided by many different exchange facilities, primarily those which describe their user and business policies.

Top of page

Packet Clearing House
Presidio of San Francisco
572-B Ruger, Box 29920
San Francisco, CA 94129-0920 USA
Tel: +1 415 831 3100
Fax: +1 415 831 3101
info@pch.net