I think I need a 100G DTN!
EPOC advice on evaluating if this is a good idea or not

EPOC Contact Point: Jason Zurawski (zurawski@es.net)
Last Edit: December 4, 2020

The Request

As part of the EPOC consultation process, we are often contacted for advice in building 100G-capable data transfer hardware (DTNs). Our response generally begins by suggesting that the folks take a step back and first identify their use cases versus making a choice based on technology availability. One mis-step that many sites make when deploying DTNs is being tempted to build the biggest they can, with the belief that not only is bigger always better, but that getting the biggest DTN possible now will mean that this approach will scale for longer. However, our experience has shown this approach is often counterproductive for several reasons that are apparent when you examine the typical requirements for data transfers of the science drivers, the available network and infrastructure around the DTN, and the additional cost factors.

Why 100G Isn't the Right Solution

Typical science use cases are not yet scaling to 100Gbps single stream TCP requirements for a number of reasons. Beyond a few limited experiments associated with high energy physics, we have not seen single-host applications able to achieve anywhere close to 100Gbs throughput across a network, generally due to other mitigating factors associated with the application workflow. For example, instruments that are capable of streaming multiple Gbps of data often must write in parallel to archival storage or be subject to a set of computational-bound filtering steps to pre-process the data. We have found that the most common use case is multiple 1-10Gbps parallel streams produced after processing that are more effectively handled by multiple smaller DTNs. Using multiple 10 or 40G DTNs enables the needed parallelization much more effectively and for lower cost, without sacrificing overall performance expectations.

For a 100G DTN to function at peak efficiency, it will need to talk to other 100G devices, which today are limited in number. Many, if not all, of major computational facilities are not operating 100G capable devices, nor do they have plans to update to this type of equipment in the near future. Some may feature 40G capable devices, but most are still operating with multiple instantiations functioning at 10G. This is in part because, operationally, most facilities do not yet have the backend cyberinfrastructure (storage, computational infrastructure, etc.) that can effectively deal with writing or reading files at the higher speeds. With few if any national-level sites aiming for native 100G capabilities, it doesn't yet make much sense for individual campuses to make such a significant investment for no gain.

In addition, it is currently very uncommon for a campus backbone to be able to support 100G. However, even with a 100G backbone, it is likely that there would be only a single user or use case driving the need for a 100G DTN, in which case, that single user will use not only the full

capacity of the DTN but also the full capacity of the network, and obviously unfeasible situation.That amount of network utilization (and potential congestion) can be destructive to the rest of the traffic on the network, regardless of whether it is acting as a sender or as a receiver.

Many regional, national, and international networks are built on 100Gbps backbones, and in some cases may even have multiple 100Gbps backbones between high traffic locations. However, this infrastructure is shared between all of their connected institutions and the likelihood of a path having near 100Gbps of available bandwidth is low. The most common network protocols will throttle traffic and try to load balance all of the users traffic.

Moreover, at least at this time, the cost of the associated 100Gbps capabilities is still very high. The optics and interfaces on both the server and the associated routers or switches will also drive the cost up significantly versus their more commercially available 10G and 40G counterparts.

Our Current Recommendations

In general, when approached with a request for 100G DTN configurations, instead we recommend that the campus or site should consider designing hardware that can scale, but not necessarily starting at the 100G level. We recommend purchasing components that can grow as required, while meeting the use case demands. We recommend:

- Purchase CPUs that are high clock rate (3.6Ghz or above) with a moderate core count. For example, 8-12 cores often works well to handle multiple data streams along with other aspects of system operation. A fast CPU is critical to the effective use of TCP and other emerging protocols. Dual processors may help depending on the expected scalability requirements of the host, but with them comes additional tuning and balancing that may complicate the set up unnecessarily.
- Maximize the use of the RAM slots when applicable. This doesn't mean getting the largest capacity for each DIMM, but it does mean filling all available RAM sockets with an equal sized memory allocation. The equal size is important as it greatly helps the machine to balance memory load. You will achieve an overall better performance baseline, even with a smaller capacity, provided that is evenly distributed between the sockets.
- Utilize an equal number of similar sized drives. Similar to approach for RAM, this approach will more effectively distribute the load during reads and writes. For example, we've had more consistent experiences when using Intel's Virtual Raid on CPU (VROC) when the system can allocate memory equally across a uniform number of devices of equal size.
- Use a network card (or cards) that facilitate swappable optics and multiple Ethernet standards. Many cards on the market today support a standard QSFP28 port, which facilitates use of compatible active cables or optics that support a variety of standards (e.g. 10/40/50/100G).

Scalability can be accomplished incrementally following this approach. It is possible to upgrade the machine directly to support larger speeds as the surrounding local network is upgraded. For example, if a campus backbone is updated to support 100Gbps operation, it is straight forward to update a 10G DTN to a 40G DTN. It is also possible to scale “horizontally”, by adding more same-speed devices, and allocating different science use cases to these to separate security profiles or duty cycles. If you use Globus, you can configure resources to be viewed as a single endpoint, and thus have additional load balancing at the application layer. If you decide to operate containers or other software sharing schemes, you can split shared resources in an effective way as well: either on the same machine or on different physical hardware. This has been extremely helpful at several sites we’ve worked with.

In general, we always recommend that the engineers thoroughly understand the use cases and build solutions to those needs. There’s no point in building something that no one will be able to use and then bemoaning that fact later on.

For more information on how ESnet and others have designed their DTNs, we recommend:

- August 28, 2020, CI Engineering Talk on DTN Design: https://youtu.be/tsNgqg27MGk as well as the accompanying slides: \https://drive.google.com/file/d/1Sv0i_Q6VIHgPyvgfeuSC0RSYDQWnQWZk/view?usp=sharing
- August 14, 2020, CI Engineering talk on the “Data Mobility Exhibition”: https://youtu.be/CmHGu9cG0ww and a related website: https://fasterdata.es.net/performance-testing/2019-2020-data-mobility-workshop-and-exhibition/
- ESnet DTN Reference Architecture: https://fasterdata.es.net/science-dmz/DTN/reference-implementation/