Using NVMe Gen3 PCIe SSD Cards in High-density Servers for High-performance Big Data Transfer Over Multiple Network Channels

02/07/2015

Chin Fang

Abstract -- This Technical Note describes how the Zettar team came up with a data transfer cluster design that convincingly proved the feasibility of using high-density servers for high-performance Big Data transfers. It then outlines the tests, operations, and observations that address a potential over-heating concern regarding the use of [Non-Volatile Memory Host Controller Interface Specification \(NVMHCI aka NVM Express or NVMe\)](#) Gen 3 [PCIe](#) SSD cards in high-density servers. Finally, it points out the possibility of developing a new generation of high-performance [Science DMZ](#) data transfer system for the data-intensive research community and commercial enterprises.

SLAC National Accelerator Laboratory, Stanford University, Stanford, CA 94309

This material is based upon work supported by the U.S. Department of Energy,
Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-76SF00515
and Citrix Systems, Inc.

Using NVMe Gen3 PCIe SSD cards in high-density servers for high-performance Big Data transfer over multiple network channels

Chin Fang, Ph.D. Founder, Zettar Inc. <fangchin at zettar.com>

Feb. 7, 2015

Table of Contents

Introduction.....	2
The current state-of-the-art Data Transfer Node (DTN) design.....	2
A key to high-speed data transfers - high performance storage.....	3
A brief introduction to the Zettar software.....	5
Design of the SC14 live demo setup.....	6
Background.....	6
Design constraints.....	6
Space and power.....	6
Hardware availability.....	6
Logistics.....	7
Schedule.....	7
Man power.....	7
Hardware configuration.....	7
Software environment.....	7
Site policies.....	7
The resulting design.....	7
A major concern - heat!.....	9
Experience and systematic investigation.....	10
Experience.....	10
Systematic investigation.....	11
Baseline information.....	11
The maximum allowable operating temperature of Intel P3700.....	11
How Intel NVMe P3700 operating temperatures are monitored and tracked.....	12
Test methodologies.....	12
Test data sets.....	12
Compression/un-compression tests.....	12
Data transfer tests.....	12
How tests are performed.....	12
Test categories and results.....	13
Compression/uncompression tests.....	13
Dark Sky Simulation single big file.....	13
Large file data transfer tests.....	14
Lots of Small Files (LOSF) data transfer tests.....	14
Conclusions.....	15
Acknowledgements.....	16

Introduction

Zettar Inc. sponsored Stanford University's [Supercomputing 2014 \(SC14\)](#) effort by live-demoing during the event a next-generation, ultra-high performance parallel Big Data transfer software, freshly implemented in 2014. The software, called Zettar $z\mathbf{x}$, is the base of Zettar Inc.'s commercial close-source data transfer solution for data-intensive enterprises.

During the intense preparation phase, the actual event, and post-SC14 follow-up effort, the Zettar team gained valuable experience in various previously unexplored high-performance and high-end enterprise computing areas. For example, how best to practically exploit the "divide and conquer" approach - a widely used key HPC principle - for Big Data transfers at high-speed, i.e. at rates $\geq 10\text{Gbps}$.

The intention of this series of SLAC Technical Notes is to share such experience with the data-intensive science community. This particular note is the first of the planned series.

The focus of this Technical Note is to describe:

1. How the Zettar team came up with a data transfer cluster design that convincingly proved the feasibility of using high-density servers for high-performance Big Data transfers.
2. The tests, operations, and observations that address a potential over-heating concern regarding the use of [Non-Volatile Memory Host Controller Interface Specification \(NVMHCI aka NVM Express or NVMe\)](#) Gen 3 [PCIe](#) SSD cards in high-density servers.
3. The possibility of developing a new generation of high-performance [Science DMZ](#) data transfer system for the data-intensive research community and commercial enterprises.

Before getting into further discussions of the above three aspects, some background information first:

The current state-of-the-art Data Transfer Node (DTN) design

The [Energy Science Network, aka ESnet](#) publishes a DTN reference design at [Data Transfer Node Reference Implementation](#). Many in the data-intensive science community consider it to be the state-of-the-art. It is similar to the following Caltech design:

The image is a promotional graphic for a "40GE Server Design Kit" from the University of California, Irvine (UCI) and the US LHCNet project. It features a dark blue background with white and yellow text. On the right side, there is a collage of hardware components including a server rack unit, a network interface card (NIC), a fiber optic cable, a SATA RAID controller, a SSD, and a network switch. The text on the left lists the components of the kit with checkmarks, and a black box at the bottom states the server cost.

40GE Server Design Kit

- ✓ **SandyBridge E5 Based Servers:**
(SuperMicro X9DRi-F or Dell R720)
Intel E5-2670 with C1 or C2 Stepping
128GB of DDR3 1600MHz RAM
- ✓ **Mellanox VPI CX-3 PCIe Gen3 NIC**
- ✓ **Dell / Mellanox QSFP Active Fiber Cables**
- ✓ **LSI 9265-8i, 8 port SATA 6G RAID Controller**
- ✓ **OCZ Vertex 3 SSD, 6Gb/s**
(preferably enterprise disks like Deneva 2)
- ✓ **Dell – Force10; Z9000 40GE Switch**

Server Cost = ~ \$15k

Logos for UCI and US LHCNet are in the top corners.

FIG. 1 A DTN design published by Caltech.

Such designs are proven and usable up to 40Gbps with just a 2U commodity server, although the aforementioned ESnet reference design Web page recommends a 4U form factor for expandability. Nevertheless, before adopting such designs, it is recommended to consider carefully factors such as space, power consumption, cost-effectiveness, and design longevity. Note also that

- A significant portion of the room inside of such designs are occupied by storage devices such as HDDs and/or SSDs, which usually are arranged as a RAID
- DTNs usually are deployed in a [Science DMZ](#)

A key to high-speed data transfers - high performance storage

Even in 2013, it was already plainly clear to both Stanford Research Computing and Zettar that high-performance Big Data transfers are no longer *network alone* tasks over high-speed network connections, i.e. 10Gbps and faster. The following three pictures, including a summary of a study by Dr. Phil Reese and the author in 2013, together with a schematic of a NASA team's [Supercomputing 2013 \(SC13\)](#) 91Gbps disk-to-disk data transfer over an 100Gbps WAN demo, should serve to illustrate this point clearly:

FIG. 2 NVMe and SATA SSDs' sequential read throughput; the respective speed of eSATA, USB 3.0, and Thunderbolt interfaces; various network speeds, where [1 MiB = 1,048,576 bytes](#). It's evident that for the Thunderbolt interface and a network speed at 10Gbps and faster, even the combined sequential read throughput of **two Intel 730 SSDs** can't saturate a single 10Gbps network link. A brand new high-end notebook, even with a SATA SSD, can attain at best 300 MiB/s or so throughput when connected to a 10Gbps link - bandwidth is not an issue in this case. It's storage I/O foremost, followed by computing power availability, and then the available bandwidth. *Note also that when layered on top with a file system, the reported SSD speeds are not fully attainable due to file system (software) overheads* - the subject of a subsequent SLAC Technical Notes. Finally, it's a must to employ software that can leverage such computing resources optimally to achieve the desired high Big Data transfer throughput.

Fast Data Transfers Demand More Than Networks - I

All 7 pairs are 10GbE connected, why the throughput differences?

Data source: Dr. Reese, Stanford Research Computing and Dr. Chin Fang, Zettar Inc. 2013

Copyright © Zettar Inc. 2013 - 2015

FIG. 3 In the Fall of 2013, Dr. Phil Reese and the author carried out a data transfer study using seven different pairs of servers of different capabilities and configurations. This study yielded seven different test results. All server pairs were connected to 10Gbps Stanford University campus LAN. All tests were conducted with two data transfer tools, Zettar's then prototype *zcloud*, and *bbcp*. A main reason of the observed differences? Storage I/O performance differences. Even the best results obtained were less than half of the network bandwidth (e.g. 400MiB/s vs. 1250MiB/s)

Fast Data Transfers Demand More Than Networks - II

Copyright © Zettar Inc. 2013 - 2015

FIG. 4 Shortly after Dr. Reese and the author carried out the study depicted in **FIG. 3**, a NASA team demonstrated at [SC13](#) a 91Gbps disk-to-disk transfer over 100Gbps WAN. A key to the achievement? Custom-made RAID units fully populated with SSDs, instead of conventional mechanical HDDs. Furthermore, [only large files are used to achieve streaming](#), which is usually much easier to yield high throughput.

A brief introduction to the Zettar software

The software that Zettar live-demoed during SC14 had its design in gestation since mid-2013. A technical talk about the software is scheduled to be given at SLAC on March 4th, 2015. The talk covers, among others, the following:

- The motivation of the design
- Unique capabilities
- Architecture overview
- Use cases

As such, this Technical Note only provides sufficient information for context. The design of the software is motivated by the following observations of the author:

1. Large enterprises often use multiple high-speed (i.e. ≥ 10 Gbps) private networks to connect their facilities.
2. Such network links are often of Ethernet style, even having the IP addresses of both ends in the same subnet.
3. None of today's data transfer solutions can use these multiple connections at the same time - for redundancy or fall-over maybe, but not for full utilization.
4. Big Data is ever growing, so scalable throughput alone is insufficient. A practically "boundlessly" scalable data transfer throughput is a must. See the following figure:

FIG. 5 IDC's Digital Universe report made it clear the last point mentioned by the author

As a result, the author created the following conceptual design:

ZX – Boundlessly Scalable Data Transfer Throughput

Copyright © Zettar Inc. 2013 - 2015

FIG. 6 The Zettar software *zx*, with sufficient backend storage I/O, enough computing power, can attain 1X data transfer throughput with a single network link connecting the two sites, 2X when there are two network links, 3X when there are three, and so on. Thus, in practice, it offers "boundlessly" scalable data transfer throughput. Note also that instead of relying on RAID, which is a host-oriented approach, the author adopted a distributed storage created using a high-performance parallel file system for superior aggregated storage I/O. See also the section **A key to high-speed data transfers - high performance storage**. Also note that the author used a cluster of compute nodes to gain high aggregated computing power, and the employment of multiple network channels for network bandwidth scale-out.

Design of the SC14 live demo setup

Described below are the background of the design and the constraints faced by the author.

Background

Stanford Research Computing started discussing with the author about having Zettar sponsor Stanford's SC14 endeavor in April, 2014. The author was informed that the main Stanford theme at the event would be high-speed data transfers, likely over the anticipated 100Gbps Stanford WAN connection to be introduced in late 2014, *right before the SC14 event!*

Design constraints

Described below are major constraints. It is of interest to note that more than half of them are [layer 8 - 10](#) in nature, i.e. user, organization, and regulation related.

Space and power

In the planned Stanford SC14 booth, only a half-rack is available, and only 1/2 of it was available for Zettar's use.

Hardware availability

Owing to budgetary constraints, both Stanford University Research Computing and Zettar decided to seek system vendor sponsorship to obtain the desired servers, switches, and other components for the event. Eventually, [QCT \(Quanta Cloud Technology\)](#) and Intel Corp's SSD division (aka Non-Volatile Memory Solutions Group, otherwise known as NSG) graciously provided the desired hardware evaluation units.

Logistics

The setup must be reliable and relatively portable at the same time, *even as a cluster*. During the staging phase, the SC14 event, and post SC14, the setup needs to be moved several times from one location to another.

Schedule

Purchasing paper work delay cut into the hardware availability for about a month, seriously impacting Zettar's preparation - instead of having 3 months to get the new software/hardware combinations working correctly, there were only two months left - with many other urgent tasks competing for the severely limited time.

Man power

Zettar had only three engineers, including the author. Being the only one in the San Francisco Bay area in a distributed team, the author was left with the challenge of completing all tasks that required the physical presence of an engineer, while tending to the numerous other tasks that a startup founder typically has to deal with - *including fund-raising!* Note that this was in addition to the personal commitments that come with being a husband and father.

Hardware configuration

As an attempt to extract as much data transfer performance as possible from the available hardware, the [OCT S810-X52L 2U 4Node high-density servers](#), as specified by the author, all came with a very unusual configuration. In addition, two of them were installed with newly introduced Intel P3700 NVMe SSDs that the author had no prior experience with. As a result, it took a while for the author to get the systems configured correctly, delaying the software development progress. Furthermore, by the time it was the most critical to do software verification, all systems were packed up and shipped out to New Orleans, Louisiana. The Zettar team had to conduct testing, debugging, and verification using an environment that was far less than ideal - a factor that significantly hampered what could have been achieved during SC14.

Software environment

Intel NSG recommended the author to target the then beta [Red Hat Enterprise Linux \(aka RHEL\) 7](#) due to the availability of a much improved NVMe driver (designed and implemented by an Intel engineer). As such, the Zettar team needed to deal with not only new hardware, but also a new operating system.

Site policies

In all data center facilities available to staging the SC14 live-demo setup, the IP address allocation, the DNS setup, firewall policies are all beyond the author's control. Thus, the minimization or elimination of interactions with each staging site's IP addressing scheme, DNS setup, and firewall setups is a main design goal.

The resulting design

Engineering is the art and science of staying within given constraints but using the right compromises to attain the most optimal technical result to support a desired business objective. After much deliberation, taking into account of all the above constraints, Zettar's unique capabilities, and the fact that a Supercomputing event is where one pushes the technology boundaries as much as one can, the author came up the conceptual design illustrated in **FIG. 6**, with the following main design elements:

1. High-density servers are used to gain space and energy efficiency, together with a low overall cost, trading away system expandability for typical DTN design components such as a hardware RAID.
2. As a compensation to the loss of a built-in RAID, the newly available [Intel P3700 NVMe Gen3 PCIe SSD](#) cards are employed to gain superlative storage I/O performance per node. See also **FIG. 2**.
3. Furthermore, since a traditional RAID controller cannot aggregate the storage I/O throughput from multiple nodes, a parallel file system is employed instead for storage I/O and capacity aggregation. The challenge then became, among the plethora of parallel file system choices, such as [Intel's Lustre](#), [Quantum's StorNext](#),

and [Red Hat's Storage Server \(aka GlusterFS\)](#), which one to choose? Based on the successful experience that [Stanford CEES](#) had had then with [BeeGFS](#), the author decided to use it for the SC14 demo.

4. Each node of a high-density server runs both as data transfer node and parallel file system storage server and client, to reduce the need of additional hardware - trading away short-term high-availability and some peak performance potential.
5. [10GbaseT port type \(copper\)](#) is selected to take advantage of the low-cost of CAT6 patch cables and the ease of using cable-colors for correct cabling in a high port-density deployment situation, trading away the much lower power consumption of 10GbE with SPF+ port type (but far more expensive optics).
6. The two QCT [T3040-LY3](#) 10GBaseT switches are used exclusively as a logic collection of L2-only switches with 4 VLANs configured in each, simplifying switch configuration and emulating the end points of a typical private [MPLS](#) network link at the same time.
7. A single bastion host is the only one that needs a single public-facing network interface that requires DNS configuration. Accessing all other nodes in the setup can be transparently achieved via a recursively tunneled `ssh jump` host technique that Zettar has been successfully using for years. The trade-off here is obviously that the entire setup has a clear single point of failure. The result is shown below:

FIG. 7 The back-side of the SC14 half-rack, the Zettar live-demo employed two QCT S810-X52L 2U 4Node high-density servers and two QCT T3040-LY3 10GBaseT switches to simulate two different facilities connected via 3 10Gbps private network links.

FIG. 8 The same setup has been hosted in SLAC Building 50 since Dec. 19, 2014

A major concern - heat!

The [Intel® Solid-State Drive DC P3700 Series Specifications](#) states that the operating temperature for AIC is 0 - 55 deg C ambient with specific air flow. The author was aware of the possibility that the back-side temperature of each high-density server node could get as high as 60 deg. C. So, a major concern about the design is potential overheating. The following two figures should make the situation clear.

FIG. 9 The bottom system shown in **FIG. 8** had two of its four nodes completely pulled out

FIG. 10 The two pulled out nodes are placed on an anti-static mat on a crash cart

FIG. 10 shows clearly that with a dual 10GbE mezz card (*left back-side in each node*) together with the Intel P3700 NVMe Gen3 PCIe SSD card (*the right back-side of each node*) in, the inside of each high-density node is really "packed". The above figure should make the concern about node overheating understandable. The question then became "how likely the overheating is to occur in a real data center environment?" The author decided to find out a real-world answer after all Intel P3700 SSD cards got installed in the two high-density systems on Sep. 05, 2014.

Experience and systematic investigation

Experience

On Sep. 11, 2014, the SC14 related hardware was moved from the CEES lab in Mitchell Building on the Stanford University main campus to [SRCE](#) on SLAC National Accelerator Laboratory campus. Newly commissioned in 2014,

SRCF is a green data center that relies on nothing more than outside air for cooling. Thus, it is an ideal place to test the overheating issue. To enable effective monitoring, the author setup [collectd](#) on each node to collect run-time statistics, including component temperature. To help with monitoring and visualization of such statistics, he used [graphite](#). The air temperature inside of SRCF is quite correlated with the outside ambient air temperature. As such, even at idle, these Intel P3700 NVMe Gen3 PCIe SSD cards reported different temperature on different days. Throughout the fall and early winter 2014, the operating temperature of these Intel P3700 NVMe devices never showed above-normal levels, both at idle and under stress tests, which are described next.

Systematic investigation

Based on what Zettar zx software would do, the author decided to carry out three different kinds of tests:

1. Compression/uncompression tests, using both conventional compression utilities and their multi-threaded counterparts both per host and in parallel (so as to increase the overall system operating temperature). The reason to carry out such tests is that performing data compression for data in-transit when the data types are appropriate seems to be a commonly requested feature.
2. Large data transfer tests (repeated)
3. Lots of small file (LOSF) tests (repeated)

The last one is the most I/O taxing. Furthermore, Zettar zx software, together with the test setup, already has the capability to give these NVMe devices a strenuous work-out in a manner rarely seen elsewhere. Thus it is of interest to observe the NVMe device temperature profiles generated by such tests. In passing the author would like to point out that at the peak of a large file data transfer, the current Zettar zx, over 3 x 10Gbps *independent* network links, often attains *more than* 22Gbps (i.e. 2.75GiB/s) throughput *without extensive OS, filesystem, and networking tuning*.

Since Intel NSG provided eight 800GB P3700s, with four in each QCT high-density 2U 4Node server, the aggregated storage pool for each data transfer cluster (of four node each) has sufficient capacity to hold large test data sets. As a result, the Zettar team routinely uses test data sets of sizes ranging anywhere from 200GiB to TiB. Representative results are reported and discussed in the appropriate sections below.

Baseline information

The maximum allowable operating temperature of Intel P3700

With [the isdct utility provided by Intel](#), the author determined that the maximum operating temperature is actually 85 deg. C. The 55 deg. C published on the Intel's Web site is likely to include a good safety margin:

```
[root@sc2u0n0 ~]# isdct help |more
isdct Utility:
```

The Intel(R) SSD Data Center Tool exposes all available management features of Intel SSDs.

The command line syntax follows the DMTF SM CLP standard (DSP0214) with the exception of the target portion of the command. The modified syntax used utilizes key properties of the target without requiring a syntactically correct CIM object path.

```
Usage: isdct <verb>[<options>][<targets>][<properties>]
```

Commands:

Help:

Retrieve a list of commands supported by Intel SSD Data Center Tool. Filter the results by providing a specific verb or command name. By default, if the filter is specific enough to indicate a single command, detailed help will be displayed

for that command.

```
help [-help|-h] [CommandName = (name)] [verb = (verb)]
```

intelssd:

Get a list of Intel SSDs attached to a system. The default behavior is to list a few key properties for each device; the options can be used to expand/restrict the output. If more than one device is installed output can be restricted to a single device by supplying its index.

[...]

How Intel NVMe P3700 operating temperatures are monitored and tracked

As alluded in the earlier **Experience** sub-section, the author setup a `collectd` based system for gathering system run-time statistics, in conjunction with the `graphite` system for storing and visualizing the collected results. To monitor, track, and visualize the operating temperature of Intel NVMe devices, the author created a [collectd Exec plugin](#) wrapping around the aforementioned Intel `isdct` utility. The resulting system is simple and easy to use. Selected results and pictures are given in the appropriate sections below.

Test methodologies

Described below are, first, test data sets employed, second, how tests are conducted.

Test data sets

Compression/un-compression tests

Two SLAC research projects are tapped as data sources:

1. **Lots of small files (LOSF):** [SLAC Pinger Ping end-to-end reporting project](#) data. The history of this data set goes all the way back to 1998. Total test data set size: 54GiB.
2. **Big file:** [SLAC Dark Sky Simulation](#) - a single 257GiB file is employed.

Data transfer tests

The data file size distributions from the [Optimizing Large Data Transfers over 100Gbps Wide Area Networks](#), section IV, **Performance Evaluation**, are adopted. Zettar however uses much larger test data sets that are programmatically generated using a utility implemented by the Zettar team:

1. **LOSF:** 391 GiB (vs. 8MiB as used by the cited paper)
2. **Big file:** 200 GiB (vs. 14GiB as used by the cited paper)

How tests are performed

All tests were conducted on nodes accessed remotely via `ssh` by the author on his [Dell Studio XPS 16 1640 notebook](#) running Ubuntu 12.04LTS amd64.

For the compression/uncompression tests, the author either did a desired test in the shell on a particular node, or used a `bash` shell wrapper of the [ansible](#) utility that in turn called the desired compression/uncompression utility to do a desired test on all nodes of a QCT 2U 4Node high-density server in parallel. Then the author observed and recorded the NVMe temperatures as reported by the `graphite` Web UI. Furthermore, the following utilities are employed as well:

1. [pigz](#) - A parallel implementation of `gzip` for modern multi-processor, multi-core machines
2. [pbzip2](#) - Parallel BZIP2 (PBZIP2) Data Compression Software

For the data transfer test, the author used a Zettar implemented Python CLI tool that calls the RESTful control APIs of Zettar zx software to carry out a desired test repeatedly. Then the author observed the NVMe temperatures as reported by the graphite Web UI.

Test categories and results

Compression/uncompression tests

Dark Sky Simulation single big file

Utility	Action	Wall-clock time	Size
pigz -9	compression	38:03.71	196G
unpigz	decompression	43:01.45	257G
pigz -9	compression	38:22.81	196G
unpigz	decompression	42:40.66	257G
pigz -9	compression	38:12.34	196G
unpigz	decompression	42:59.62	257G
pbzip2 -9	compression	1:01:56	193G
pbzip2 -d	decompression	27:03.00	257G
pbzip2 -9	compression	1:01:43	193G
pbzip2 -d	decompression	27:06.12	257G
pbzip2 -9	compression	1:01:58	193G
pbzip2 -d	decompression	27:04.91	257G

FIG. 11 `pigz -9` may be faster than `pbzip2 -9` in compression, but when it comes to decompression, the table turned 180 degrees. Note also that during the test, the employed NVMe device's operating temperature only went up about 5. deg. C. The above test was conducted while the hardware was hosted in the SRCF prior to SC14.

Since the author did not observe any significant operating temperature differences during other tests, including parallel compression/uncompression tests, for the sake of brevity the results are not reported here.

Large file data transfer tests

Zettar *zx* is a cluster software by design. Thus, when conducting a data transfer operation, all nodes in a data transfer cluster are employed. Consequently, all NVMe devices in such nodes are utilized as well. It should be evident from the following graphite display that during the testing, even on the write-intensive side, the NVMe devices' operating temperature rose by less than 10 deg. C.

FIG. 12 Intel P3700 operating temperature during large file data transfer testing. Note that the NVMe devices in the bottom system were having intensive writes, and thus their temperature profiles differed from their counterparts of the read-intensive NVMe devices in the top system. The bottom system also must carry out data reconstitution by combining numerous data chunks into the original Big Data file. The above test was conducted while the hardware are hosted in the SLAC data center post SC14.

Lots of Small Files (LOSF) data transfer tests

In this case, as in the case of large file transfers, over the approximately half-an-hour non-stop testing using a LOSF test data set that has a total size 391 GiB and 20000 files of sizes varying from 8KiB to 4MiB, even the sharpest operating temperature profile shows less than 10 deg. C rise.

FIG. 13 Intel P3700 operating temperature during five repeated LOSF data transfer testing. Note again that the NVMe devices in the bottom system were having intensive writes, and thus their temperature profiles differed from their counterparts of the read-intensive NVMe devices in the top system. The above test was conducted while the hardware are hosted in the SLAC data center post SC14.

Conclusions

Based on what the Zettar team has accomplished so far, the author is confident that the combination of modern high-density servers (e.g. QCT's [QuantaPlex T41SP-2U \(4-Node\)](#)), the latest NVMe SSD devices such as [Intel's P3700](#) series (see, e.g. **FIG. 14**), and Zettar $z \times$ parallel data transfer software, facilitates the construction of cost-effective, space and energy efficient, and very high-performing data transfer clusters. Specifically, given the fact that Zettar $z \times$ data transfer software is designed and implemented using an advanced distributed, highly-concurrent, and asynchronous software programming model, the author anticipates that the result will out-perform the current state-of-the-art DTN design by more than 200% *with the same form factor*, among other benefits.

FIG. 14 During SC14, Intel NSG unveiled its P3700 NVMe SSDs in both PCIe AIC and 2.5" form factors

Acknowledgements

The SC14 endeavor is truly inspiring and challenging to the Zettar team. It should be noted, however, that without the assistance and support of many helpful people and businesses, it wouldn't have been possible at all. Specifically:

Dr. Phil Reese, Stanford Research Computing, for having Zettar as part of the SC14 endeavor. Ruth Marinshaw, CTO, Stanford Research Computing, for her cheerful support and encouragement. Dennis Michael, CEES, Stanford University, for mentoring the author regarding large computing site management during his Ph.D. days, always being encouraging, and granting access to the CEES Lab during the initial SC14 staging period.

The author gratefully acknowledges [QCT \(Quanta Cloud Technology\)](#) for the loan of the [S210-X12RS](#) 1U servers, [S810-X52L](#) high-density servers, and [T3040-LY3](#) 10GbBaseT switches. Raymond Huang and Jason Pan, for initiating QCT's hardware sponsorship that made a successful SC14 showing possible, together with the friendly support from the two gentlemen and their various FAE colleagues. Intel Corp., for the loan of its superlative [P3700 NVMe Gen3 PCIe SSD cards](#). Ranjit Kurup, Long-Range Business Strategy, DCG/Storage Group, Intel Corp., for his encouragement and help putting the author in touch with various Intel product groups, leading to the aforementioned hardware loaners. James Myers, Intel NSG Manager, for his quick perception that Intel NSG's then newly introduced NVMe product would be a great match to our SC14 endeavor and generously offered the loaner units. Jack Zhang, Sr. SSD Application Engineer, Intel NSG, for the many useful suggestions during our preparation phase, and for providing the Intel SSD Data Center tool `isdct` for monitoring Intel P3700s. Andrey O. Kudryavtsev, Cindi Peach, both also of Intel NSG, for product related discussions. To Chris Summers of Arista Networks, Inc. for two [7050TX-48](#) loaner switches. They came at the right time and enabled Zettar to continue its development efforts when the actual setup was shipped out. [The Fraunhofer ITWM Competence Center for High Performance Computing](#), for making its robust, fast, and simple-to-install-and-use BeeGFS [available free to use](#). Jan Hichler of [ThinkparQ GmbH, a Fraunhofer spin-off](#), for being a pleasure to work with. Sven Breuner and his team (esp. Bernd Lietzow), of Fraunhofer ITWM, for their quality software engineering.

To the many helpful colleagues at SLAC National Accelerator Laboratory, the author specifically wishes to thank Dr. Les Cottrell, for his enthusiastic support and encouragement, both before and after the SC14 event, and for generously providing the valuable [Pinger Ping end-to-end reporting project](#) data that he has been painstakingly collecting since 1998. Antonio Ceseracciu and Dr. Yee-Ting Li, for many helpful discussions. Randy Melen, for making it possible for the author to finish his Ph.D. at Stanford University in the first place, and for his support and encouragement. Lance Nakata, for being a helpful friend since the author's Ph.D. days and for finding the right data file for the tests. Andy Hanushevsky, for being the brave sole creator of the [bbcp](#), one of the two leading DOE high-performance data transfer tools - the much easier to use one.

To a good friend Matt Petach, Yahoo! Inc., for the many stimulating conversations over the years, and for quickly bringing the author up to speed in configuring Cisco-compatible switches in the tight pre-SC14 period.

To his two Zettar colleagues, Igor Solovyov and Alex Nazarenko, for their can-do attitude, for their [software craftsmanship](#), and for pulling off a "Mission Impossible" together with the author.

The SC14 effort is partially supported by the seed funding from Citrix Systems, Inc. The author wishes to acknowledge Dr. Michael Harries, Chief Technologist, Senior Director, Citrix Startup Accelerator, for making the funding possible.

Last but not the least, the author gives his most heart-felt thanks to his family, for the unwavering support, especially during the nights the author had to work late in the data center.