CRU

Protecting
Your Data™

Removable NVMe SSD Storage: Technical Considerations

For over 30 years, CRU removable storage products and technology have been entrusted to secure sensitive data for military and government agencies and well-known organizations all around the world.

This paper will discuss the considerations the CRU engineering team has taken into account when developing robust, removable storage solutions for PCIe/NVMe SSDs. To develop a dependable removable solution for a high-speed storage technology that wasn't originally designed to be removable, the team had to consider:

- Signal integrity. With high-speed signals employed to move data across multiple interconnects, maintaining the integrity of the data-carrying channels is crucial.
- Thermal issues. NVMe SSD controllers and NAND memory chips generate a large amount of heat, especially when constrained to enclosed spaces. High levels of throughput performance is one of the reasons to incorporate NVMe SSDs into compute and data storage platforms. When SSDs are too hot, their performance throttles, defeating the reason for incorporating high-speed storage into a removable system solution.
- Durability. A standard, off-the-shelf NVMe SSD employs a connector designed for break/fix replacement, not the day-in/day-out removability required by data security usage or applications that rely on sneakernet for data transport.
- Ruggedness. When packaging NVMe storage into a module designed for non-office environments, design considerations for waterproofness, shock and vibration resistance, and operation over extreme temperatures need to be taken into account.

SIGNAL INTEGRITY

THERMAL ISSUES

DURABILITY

RUGGEDNESS

SATA stalls, PCIe/NVMe races ahead

Historically, CRU has provided removable drive solutions for every storage protocol that PCs and computer workstations have implemented. Most recently, this has meant full support of SATA 3G and 6G protocols.

The SATA standard and protocol has stalled on the performance treadmill. As motherboard and backplane architecture speeds have continued to increase, SATA devices became a performance bottleneck.

Enter PCI Express (PCIe) and NVMe. The PCIe architecture, NVMe protocol, and SSDs have stepped in to go where SATA cannot. As SSD devices are closing the $/GB gap, NVMe media have become more prevalent in compute devices—due to performance as well as demands for shrinking form factors.

While these trends and transitions have been taking place, the need to secure and transport data contained on storage devices, regardless of form factor, performance characteristics, or protocol, remains a persistent requirement for government and defense users. We also see a desire for NVMe storage removability in industrial applications such as transportation (autonomous vehicle data collection, for example), agriculture (drones for imagery and sensor data collection, for example), and other industrial applications (factory automation and artificial intelligence, for example).

Typical removable NVMe storage architecture

On the left-hand side of Fig. 1, the PCIe bus originates on a motherboard, whether in a typical desktop computer or workstation or in a custom-built device. The bus is connected to the CPU or platform controller hub (PCH) and data signals traveling on the bus need to be routed to a PCIe connector or M.2 slot.

To create a removable solution, the following components are incorporated and become part of the signal transmission chain:

- Connection to the motherboard. This provides connectivity between the computer motherboard and the removable drive mechanisms. Examples include inserting a connector into the NVMe socket in place of the NVMe SSD or a host bus adapter (HBA) that is inserted into a computer PCIe slot.
- Cable. A cable capable of carrying high-speed data transmissions connects the motherboard to the receiving frame that holds the NMVe storage module when it is in place.
- Connectors. The docking connector pair used to mate the storage module to the receiving frame must be of sufficiently high quality to transmit high-speed signals—not to mention stand up to the rigors of tens of thousands of mate/demate cycles.

One way to connect to those would be to plug in a host bus adapter card (HBA) or add-in card; that card might

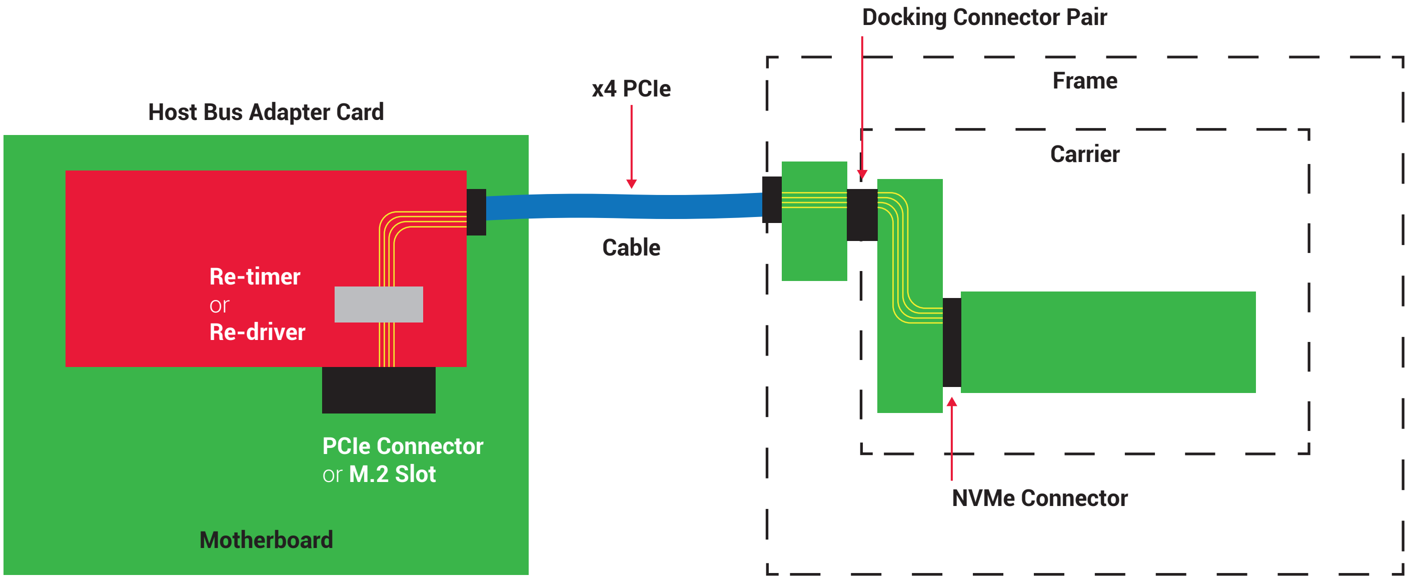


Fig. 1

have a retimer or redriver chip on it in order to condition those signals. Because we're going to be transmitting them across a cable (cable length may vary depending on the solution) at x4 PCIe bandwidth.

On the right side of Fig. 1 the cable connects to the removable device. The removable device consists of several printed circuit boards and the docking pair connector. The docking pair connector enables removability of the storage module. It is important to include a very highly durable connector that is capable of transmitting PCIe signals.

And finally we have the NVMe connector that the M.2 SSD is plugged into.

The point here is that by enabling removability, we've added nonstandard elements to the PCIe channel that aren't covered in the PCIe specification. The added cables, extra printed circuit boards, the connectors—while each may be individually rated for PCIe signal speeds, that doesn't mean that the total solution is.

Signal integrity is crucial

Conceptually, creating a removable storage device is straightforward: develop a housing for the storage media and a mechanical device to mate the media to the computer. However, with the high speed signals that PCIe devices communicate over, maintaining signal integrity becomes a crucial design consideration for a removable storage architecture.

Consider the diagram Fig. 2, which accounts for the various stages a data signal must pass through—while staying within the PCIe channel budget so the overall solution maintains PCIe compliance.

Three cases of implementation vs PCIe channel budget

The first step in achieving reliable data transmission from the compute host to the NVMe storage module is to account for the various steps the signal needs to take along its path. In desktop computers and workstations, NVMe SSDs were implemented as devices that reside directly on a computer's motherboard.

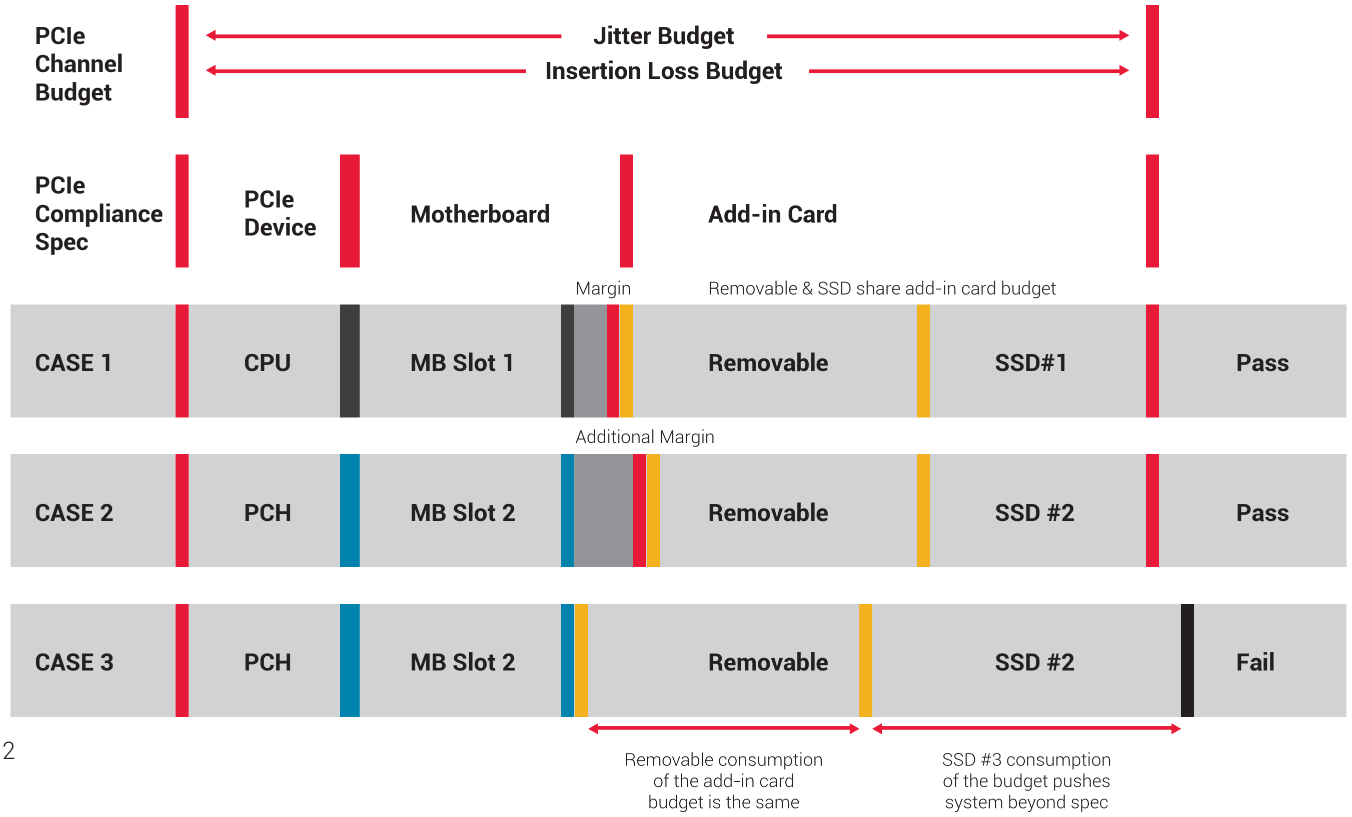


Fig. 2

PCIe channel budget

To meet PCIe compliance, the signal loss across the components used in a solution must be less than the overall budget. The PCIe channel budget is primarily made up of two elements: the jitter budget and an insertion loss budget.

An undesirable effect, jitter is created by differential signals deviating from their expected patterns. If not designed properly, the various components in a communication path (between motherboard and removable SSD in this case) can create enough jitter to render a signal path unreliable, thus disrupting the integrity of the data read or written to/from the NVMe SSD.

Insertion loss is essentially how the signal is attenuated by all of the elements in the channel, resulting in a loss of voltage or signal strength.

To meet PCIe compliance, a system design needs to consider the following contributors to signal loss:

- The PCIe device, which on the motherboard could be a CPU, a chipset (PCH), or some other PCIe device.
- The motherboard itself claims a portion of the overall budget with its routing traces and PCIe connector.
- And the final portion of the budget to account for is the add-in card. In our PCIe/NVMe removable case, the add-in card budget includes jitter and insertion loss across the host bus adapter (HBA), the removable device hardware, and the SSD media.

Three design implementation cases

Fig. 2 shows three different implementation cases and the importance of taking all factors into account when creating a reliable solution.

In the first case, the originating PCIe device is the CPU and it's connected to the motherboard slot number one. It may have some margin that comes from both jitter and insertion loss budget compared to its own budget. In this case, the removable might absorb some of that excess budget and the SSD might use the rest of the excess budget. The result is that when the various components are taken into account, the sum of jitter and insertion loss is within the overall PCIe channel budget—and the solution passes PCIe compliance requirements.

Case 2 demonstrates a different set of PCIe lanes originating from the PCH, which are connected to a different motherboard slot than in the first case. This second case illustrates that additional margin is available because the PCH motherboard may have additional jitter budget that the other slot didn't have. In this case, the removable drive device is the same and SSD 2 has a similar jitter and insertion loss to Case 1.

Case 2 is also compliant to the PCIe jitter and insertion loss budgets.

In Case 3 we're using the same PCIe slot as Case 2, as well as the same PCH and removable drive.

The removable consumption of the add-in card budget remains the same but Fig. 2 shows how SSD 3 takes up a larger portion of the add-in card budget and it is this component that has pushed the whole channel into a failing condition.

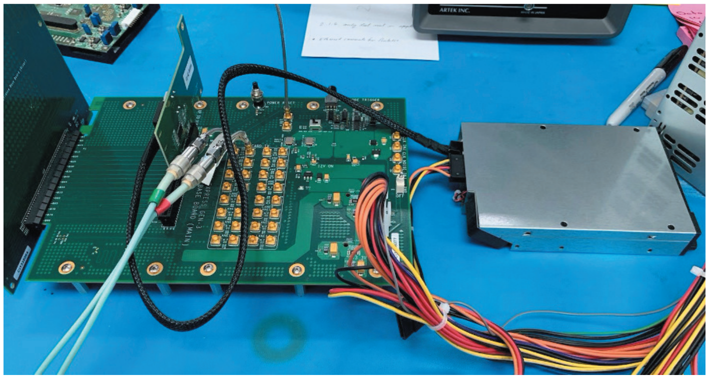
Limitations of non-compliance

The importance of staying within PCIe compliance specifications cannot be stressed enough. We've noticed through experimentation that the read/write bandwidth is limited and performance that should be at PCIe Gen 3 speeds is downgraded to PCIe Gen 1 speeds, a 3x reduction in performance.

When a component or combination of components, including the SSD itself, exceed PCIe specifications, the SSD performance can even be degraded so severely that the channel will be nonfunctional—the SSD can't even be seen by the host because the lack of signal integrity doesn't allow the SSD and host to communicate.

In our testing, we have observed that non-compliance may cause inconsistent results. Under some conditions, the SSD communicates at full speed with the host, but when another variable is introduced—such as a temperature change or a wearing down of the connection components—the PCIe bus is rendered non-functional.

While it might seem obvious, we have confirmed through testing that not every NVMe SSD behaves the same. This makes it incumbent upon the system or solution builder to ensure that SSDs and their surrounding components are thoroughly tested and selected based on performance—not price nor necessarily long-standing vendor preference.



PCIe compliance testing requires very specialized equipment, as shown in the photograph above. This photo shows the PCIe SIG compliance board that is used to test add-in cards, which is the portion of the specification that is pertinent to removable storage solutions.

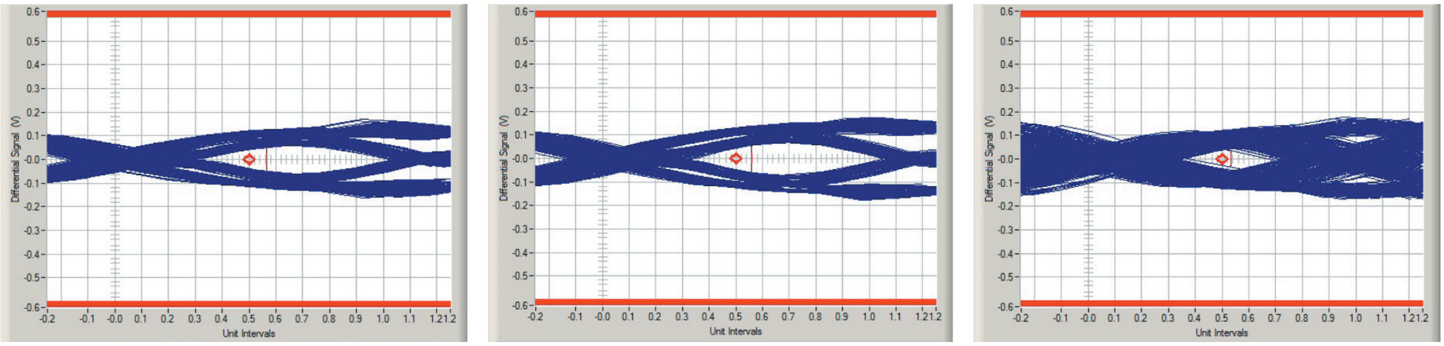
Further, we recommend using this card test method for each SSD you need to qualify for use, as well as account for all of the variables your solution might encounter: temperature, cabling, host card/bus adapter, and so forth.

Test results: transmit eye patterns

When we test our removable drive systems, we examine transmit eye patterns, which measure both signal loss and jitter. In the diagrams below, you'll see that as jitter increases, the crossover points for the differential pair get fuzzier because they're changing over time.

As insertion loss increases, the height of the eye collapses. (This collapse is related to voltage loss, as well as insertion loss and jitter.) SSD Sample 1 represents a healthy eye height. Similarly, SSD Sample 2 is a healthy height—actually a bit better with a wider eye. And SSD Sample 3, which is from the failed Case 3 above, shows a collapsed eye height—along with fuzzier crossing points caused by increased jitter.

Next are the received signal tests, where the SSD is transmitting back to the host. We use an automated oscilloscope-based test procedure. Shown are the results for Sample 2 and you can see those blue dots are within the upper and lower limit and therefore demonstrate a passing condition. Now in the case with SSD Sample 3, the received compliance test was so far out of spec that the oscilloscope couldn't run the test. As a result, we had no specific data returned that we could compare against the passing case.



SSD Sample 1
Preshoot = 3.168 dB
De-Emphasis = -5.209 dB
Vb = 217.324 mV

SSD Sample 2
Preshoot = 2.402 dB
De-Emphasis = -4.194 dB
Vb = 259.267 mV

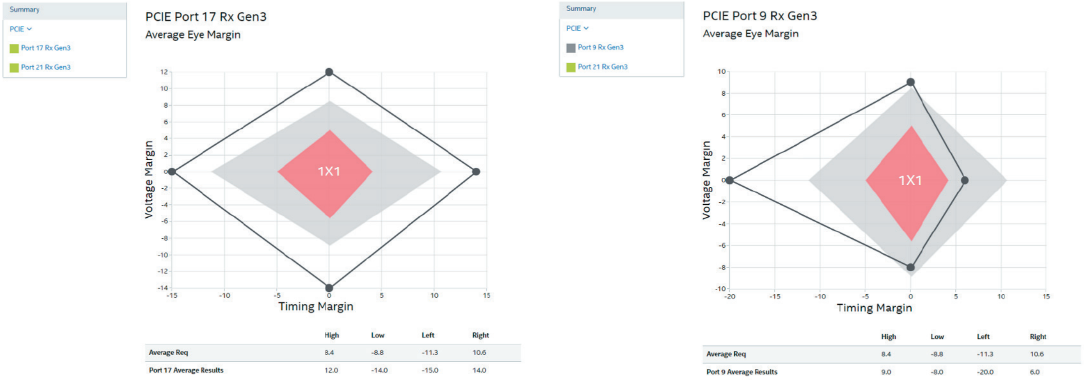
SSD Sample 3
Preshoot = 2.871 dB
De-Emphasis = -5.546 dB
Vb = 226.050 mV

Fig. 3

Intel IO Margining Tool

Another tool that we use for evaluation is the Intel IO Margining Tool (IOMT). This tool controls the PCH or CPU and changes the margins of the device to perform a loopback test. Here again in this case study we have Sample 1 with healthy margins. (The gray area in Fig. 3 is essentially the area that you want to stay out of and in the left hand case the results did so and therefore we're passing.)

But on the right, with SSD Sample 3, the eye has collapsed into the gray and therefore indicates that this was not a healthy channel.



PCIE Port 17 Rx Gen3

Average Eye Margin

	High	Low	Left	Right
Average Req	8.4	-8.8	-11.3	10.6
Port 17 Average Results	12.0	-14.0	-15.0	14.0

PCIE Port 9 Rx Gen3

Average Eye Margin

	High	Low	Left	Right
Average Req	8.4	-8.8	-11.3	10.6
Port 9 Average Results	9.0	-8.0	-20.0	6.0

Fig. 4

The Intel IO Margining tool can also be used to test the integrity of the PCIe channel. However, we do not use it as the ultimate compliance indicator. (We rely on the actual PCIe SIG method to guarantee compliance.) We use the IO margining tool as an easily-run indicator for a general idea of how well the solution is working.

Designing host bus adapter cards

Signal routing must be carefully done on these cards to match lengths of the differential pairs. In this example, the differential pairs have been very closely matched. Layer transitions must be limited to not have too many vias between layers. Vias create impedance discontinuities that can create insertion loss and add jitter. We also avoid ground planes around the signals to maintain the 80 ohm (+/- 10) impedance required by the PCIe specification.

And then also as illustrated in the architecture (see Fig. 2), redriver chips can be deployed on the host bus adapter cards to boost the PCIe lane signals and compensate for signal loss through the cables. PCIe redriver chips need to be tuned to optimize for the insertion loss margin of the channel design. Some channels can be optimized on the fly via software and some of them are specially tuned in hardware and then locked down with strapping resistors.

For more complex transmission line topologies, retimer chips can be used, which essentially recreate the PCIe bus. The PCIe bus from the host is terminated at the retimer and then recreated to make the connection to the SSD on the other side whereas redrivers are analog buffers in series with the PCIe signals.

Cable construction

We use high-quality twinax conductors in our high-speed cables to ensure that we get reliable signal quality. Here are some illustrations that show the OCuLink cable connection and the twinax cable construction. Even when these materials and cable types have been specified, they should be thoroughly tested to ensure they meet critical transmission specifications.

The right-hand side of the illustration shows return loss test results for the cable; the eye diagram measures both

jitter and insertion loss. Recall that a removable device design needs to be thought of as a complete system in and of itself.

As a result, and as demonstrated, individual components that are rated for PCIe signal speeds and SSDs that are PCIe-compliant on their own may not result in a compliant system when combined. Interchangeability of these components is limited by the lack of standards for removable memory solutions. The PCIe SIG did not consider removable storage when creating its specifications. Drive removability is a specialized type of application and even in the NVMe extensions to PCIe there have been no provisions added to the PCIe specifications.

For an NVMe removable drive architecture, system integrators must tightly control all the elements that make up the PCIe channel to ensure compliance. This helps ensure that a functional solution obtains maximum channel bandwidth, as well as consistent and reliable operation.

It is also incumbent on system integrators to perform rigorous compliance testing on all the variable components to ensure reliable performance—if the solution includes a mixture of cables from different vendors, different cable lengths, different add-in cards, and so forth, everything must be tested.

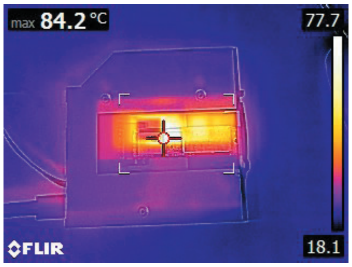


Fig. 5

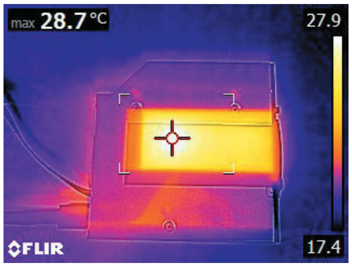


Fig. 6

Thermal management

NVMe offers significant performance but that comes at a cost: heat.

Figs.5 and 6 show the heat generated by SSDs under different workloads. The right image shows a light workload with the thermal imaging showing that the SSD is heating up to just above ambient temperature (around 25 degrees C). The left diagram shows imaging from an SSD under very heavy workload. The SSD has generated enough heat such that it is starting to throttle itself so it can still operate without shutting down completely. A typical self-throttling threshold is between 70 and 80 degrees Celsius.

To ensure that the SSD (and system) can operate at full bandwidth, some form of thermal management is necessary. CRU solutions use fans for additional airflow, in addition to temperature-monitoring circuitry, to mitigate the rise in SSD temperature.

Selecting fans, designing the ducts for airflow, and controlling fan speeds are part of the system design process. Performing thermal analysis helps the engineering team optimize the design before spending cycles to develop prototypes. Once prototype devices have been created, they are put through rigorous test cycles to obtain empirical evidence that the forthcoming products will perform as expected and desired.

Fig. 6 shows the results of a thermal simulation that shows some memory chips getting hotter than others, as shown by the darker orange color.

Observing thermal characteristics is not enough—thermal transfer must also be optimized to ensure that the SSDs can stay cool throughout operation. Most M.2 NVMe SSDs do not contain integrated heat sinks to facilitate this thermal transfer.

Thermal interface material conducts heat to the heat sink. The thermal interface material can be thermal grease or various elastomer materials. In either case, the material is applied to the top of the SSD and coupled to the device heat sink. The heat sinks and thermal interface material must be carefully designed and assembled to achieve optimal performance. Additionally, no debris can be introduced into the thermal interface material during manufacture because that will degrade its performance.

Fan performance control is also very important in the system. Fig. 7 shows a fan performance curve that illustrates the airflow rate measured in cubic feet per meter (CFM) over the static pressure of air that it is creating.

It is not enough to simply add a fan of sufficient airflow to a removable drive solution. Other requirements

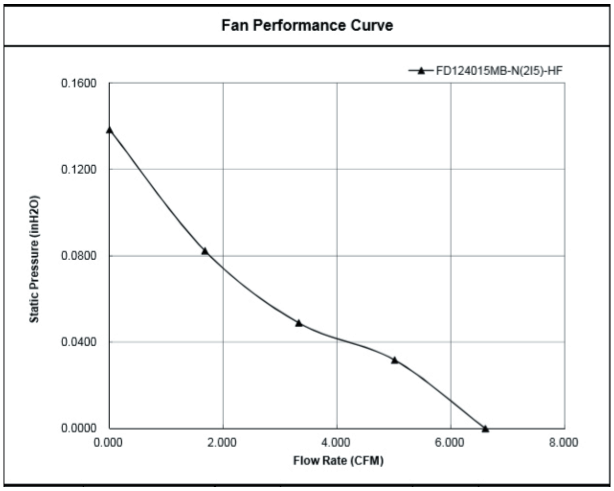
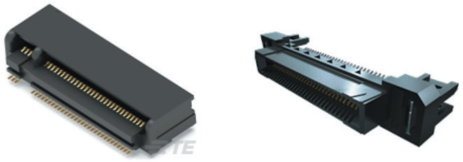


Fig. 7

such as system acoustics must be accommodated—hence the ability for a computer host system to control the removable device fan speed: high speeds when performance may be impacted and low speeds to maintain a quiet computer.

Figs. 8 and 9 contain two examples of control mechanisms that might be used. Fig. 8 shows a gradual, linear fan speed increase such that once an ambient temperature is reached on the SSD, the device microcontroller would turn on the fan and then start ramping it up to as the temperature increases.

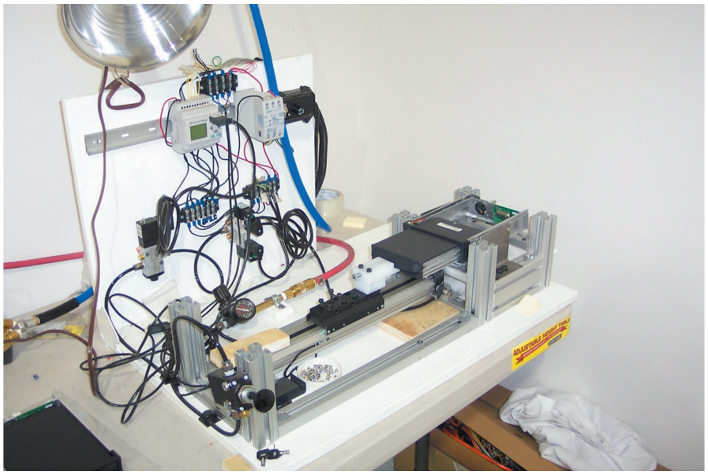
Fig. 9 shows a different, simpler scheme—a high-low scheme in which the fan might default to a slow speed of 3000 RPM, which would be barely audible. Once the temperature of the module reaches a certain point—in this case 55 degrees C—the fan would be turned to its high speed. In this case, that would be 8000 RPM.



Designing durability

CRU has a long legacy of providing ruggedness and durability for customer applications. In addition to designing solutions for the rough handling of removable drives, connector durability is critical. Without solid electrical signal paths provided by the connector, the host computer simply cannot detect, let alone read or write to, the NVMe SSD.

Consider an application that requires 10 daily cycles of removing and replacing the media—for five years. This requirement is typical and amounts to just under 20,000 mate/demate cycles.



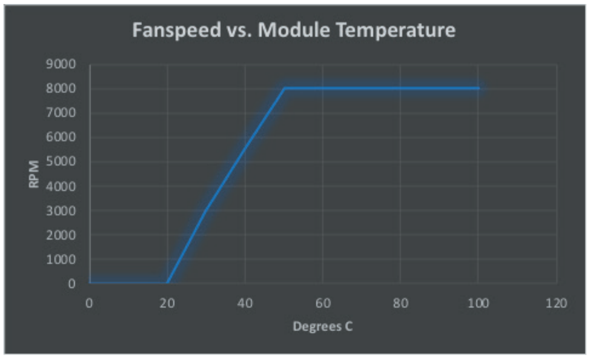


Fig. 8

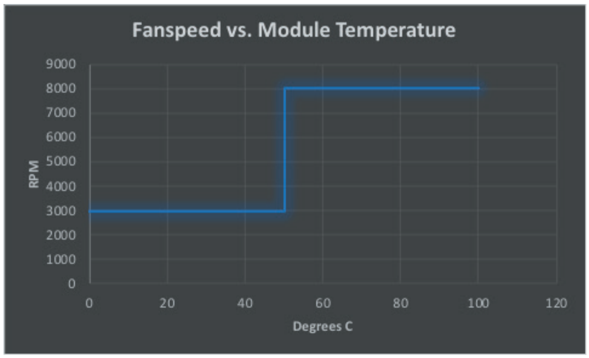


Fig. 9

A standard connector found on an unmodified M.2 SSD is only rated for fewer than 100 cycles. Relying on a standard connector clearly is not feasible most, especially high cycle, applications.

Furthermore, high-speed signal rated board-to-board connectors that are good for PCIe signaling speeds are still only rated for fewer than 1,000 cycles. We consider 20,000+ cycles to be a minimum requirement for most of our customer applications.

How can the lifespan be increased for connectors found on standard SSDs? Durability needs to be designed into the removable solution.

By controlling the mating position with guide rails and alignment features, the connector experiences less mechanical stress. The connector contacts themselves would otherwise be damaged by misalignment due to wear.

Unique insertion and extraction mechanisms can control the mate and demate forces, which preserves contact integrity. In the example shown on the right, guiderails and alignment features have been incorporated to guide the module as it mates with the circuit boards in the assembly.

As the NVMe storage module is inserted and removed, levered extractor doors were incorporated to gently extract the module by prying it out in a very repeatable fashion. The levered doors also assist in inserting a module such that the insertion force is not excessive and is also very consistent.

Again, thorough testing must be conducted to derive empirical evidence and validate the insertion and removal aspects of the design. This photo shows a fixture to measure mate and demate cycle requirements and to guarantee that we meet the specifications that we place on our product to meet customer requirements.

In conclusion, designing high-speed NVMe removable storage into a system requires very careful consideration of PCIe signal integrity, thermal management, and connector durability.

For more information,
visit the CRU web site.

cru-inc.com
sales@cru-inc.com