The Importance of Clocks in Data Centers

Hao Zheng
Systems Engineer

At a glance

1 **A closer look at data centers**
Data centers process, store and relay data through servers and switches.

2 **Clocking in data centers**
There are two types of Peripheral Component Interconnect Express (PCIe) clocking architectures that serve several purposes within data centers: common clock (CC) and independent reference clock (IR).

3 **Trend toward lower jitter**
As Ethernet data rate increases, lower jitter is necessary for high-speed SerDes.

4 **Greater integration**
BAW integration improves reliability and reduces jitter, size and cost.

Artificial intelligence (AI) and various cloud services are driving the growth of data centers. AI training requires more computing resources, while cloud services such as media streaming require more storage and data processing.

The growth in storage and computing necessitates higher connectivity speeds. As a result, the PCIe 6.0 link data rate was boosted to 64 GT/s, and Ethernet lane speeds are as fast as 224 Gbps. Higher-speed data links require lower-noise clocks to maintain high data quality.

A closer look at data centers

As shown in **Figure 1**, data centers typically comprise racks of servers. On top of each server rack is a ToR switch that relays data packets between servers and the network. A spine or fabric switch is a higher-layer switch that connects the ToR switches to the network.

High data rates usually require active cables between servers and ToR switches and optical modules between ToR switches and spine or fabric switches to reduce losses. **Figure 2** shows the server blocks that usually require clocking components.

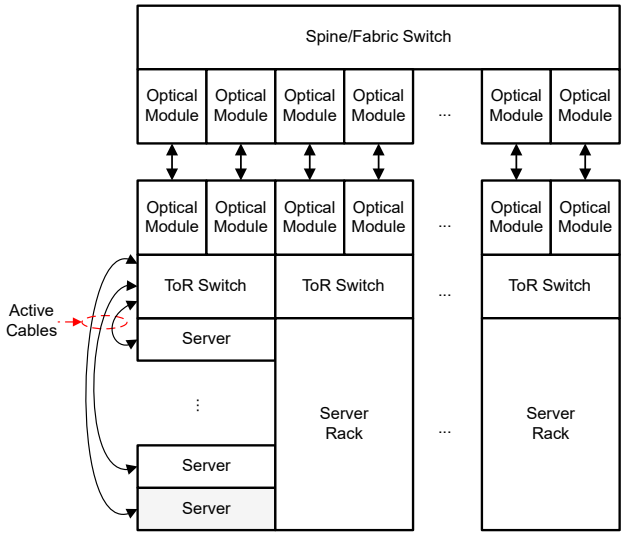

Figure 1. *Data center architecture.*

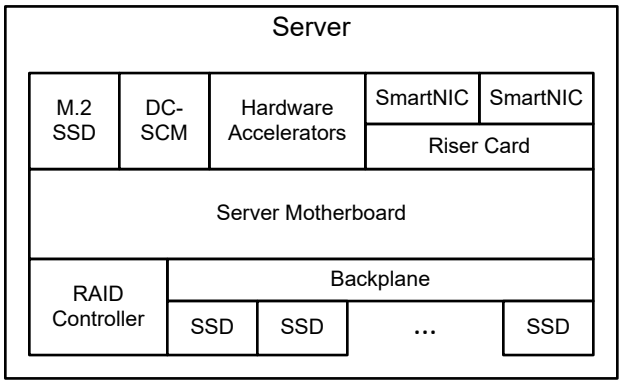

Figure 2. *Server blocks.*

Clocking in data centers

Figure 3 and **Figure 4** illustrate the PCIe internal clock and external clock architectures on server motherboards, respectively. On internal clock servers, the CPUs generate PCIe clocks. These PCIe clocks are then fanned out by PCIe clock buffers. The buffered outputs clock various endpoints or pass to daughtercards through PCIe connectors.

On external clock servers there could be various PCIe clocking sources: local PCIe clocks from the external

clock generator or PCIe clocks generated by the CPUs. Every endpoint or connector can select from one of these sources, depending on the clock domains they belong to.

Devices and interfaces usually require low-frequency single-ended clocks provided by oscillators or a clock generator.

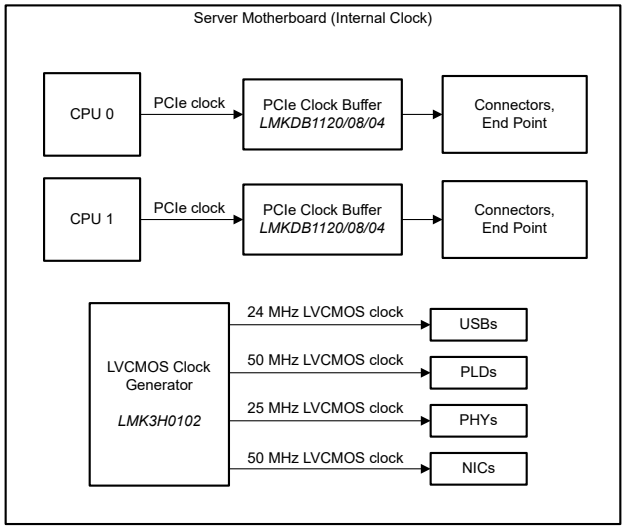

Figure 3. *Example server motherboard internal clock architecture.*

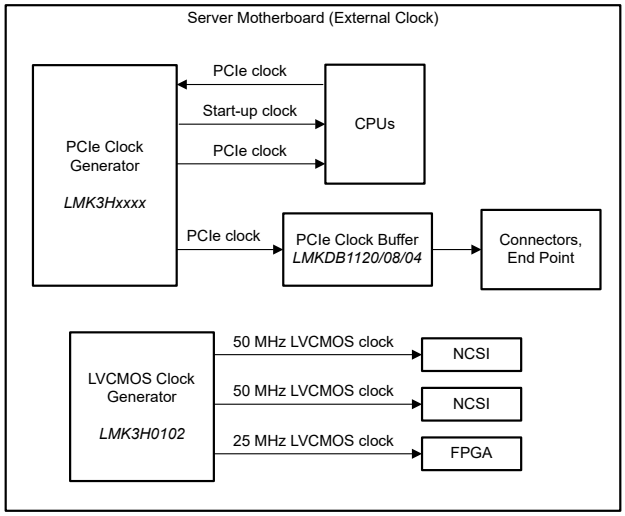

Figure 4. *Example server motherboard external clock architecture.*

A data center security control module (DC-SCM) is an add-on card defined by the Open Compute Project. In the example shown in **Figure 5**, a PCIe reference clock is provided to the DC-SCM card from the server motherboard. However, both the baseboard management controller and USB host controller need PCIe clocks. You cannot simply divide the trace and route one clock to both devices, because that would halve the amplitude and degrade signal integrity. As a result, the clock signal will no longer meet PCIe compliance, which is why a PCIe clock buffer is necessary. The clock buffer receives one clock input and generates multiple copies of the input without degrading signal integrity.

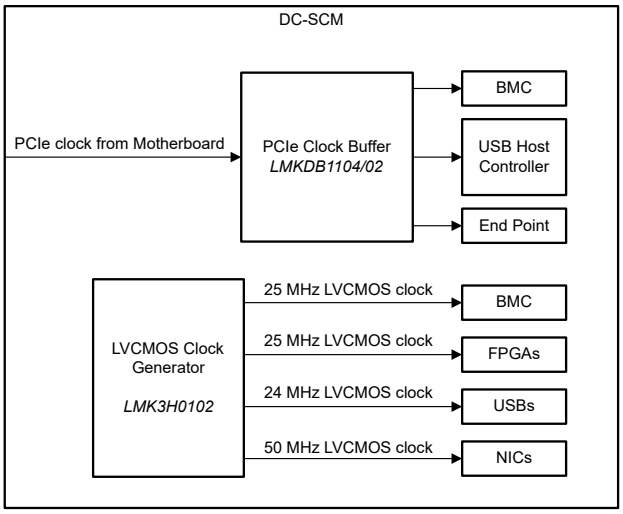

Figure 5. *Example DC-SCM clock architecture.*

Similar to a DC-SCM, other expansion cards or PCIe add-in cards may also need a clock buffer for PCIe clock distribution, as shown in **Figure 6**.

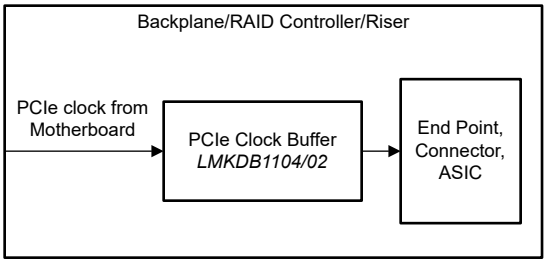

Figure 6. *Example PCIe add-in cards clock architecture.*

A network interface card (NIC) connects a server to the network. A SmartNIC provides additional computing resources to offload the server CPUs. Both NICs and SmartNICs need PCIe and Ethernet clocks. In the

example shown in **Figure 7**, there are two PCIe clock sources: a common PCIe clock from the motherboard for the CC architecture and a local PCIe clock for the IR architecture. In normal operation mode, the NIC operates on the CC. But if the CC is lost or unavailable, the NIC can switch to IR and use the local PCIe clock instead. Additionally, because a NIC connects to a switch through Ethernet ports, the Ethernet SerDes within the application-specific integrated circuit usually requires a high-performance 156.25-MHz clock.

The PCIe clocking requirements for a hardware accelerator, used for specific computing tasks such as AI training, are similar to a SmartNIC. **Figure 8** shows an example of a PCIe clock architecture. Instead of having both CC and IR architectures and switching between the two, the PCIe clocks are only generated locally. In this example, the CPUs, graphics processing units and other endpoints require many clocks. Therefore, a two-channel clock generator along with two 20-channel clock buffers can generate as many as 40 PCIe clocks. A hardware accelerator does not need an Ethernet clock because it is not connected to a ToR switch like a SmartNIC. There may be proprietary links other than PCIe that could require additional high-performance clocks, however, similar to Ethernet clocks.

Figure 9 is another example of using only an IR PCIe architecture. A two-channel PCIe clock generator is used to clock the solid-state drive (SSD) controller.

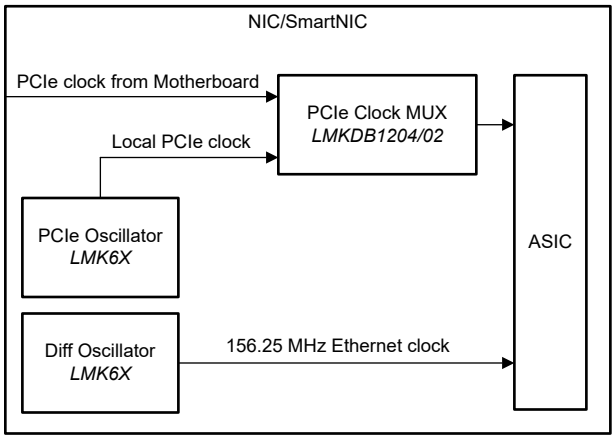

Figure 7. *Example NIC and SmartNIC clock architecture.*

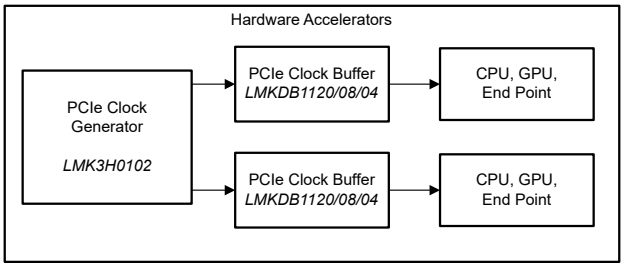

Figure 8. *Example hardware accelerator clock architecture.*

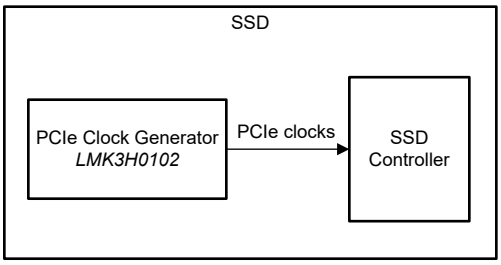

Figure 9. *Example PCIe SSD clock architecture.*

Ethernet lane speeds 56 Gbps or higher significantly affect the insertion loss of any passive cable. Therefore, “active” interconnection is necessary to reduce losses and improve data quality. Depending on the distance, there are different types of active interconnection. Active cables, including active electrical cables based on copper and active optical cables based on fiber, can connect over short distances, such as the distance from a NIC to a ToR switch.

Optical modules are used to connect over longer distances. There are also different types of optical modules. Some are used between the ToR switch and the spine or fabric switch within a data center, while others can be used between data centers.

Because of the high Ethernet lane speeds, the optical module digital signal processor requires a very low noise Ethernet clock, as shown in **Figure 10**. On the other hand, the Ethernet retimers in an active cable only need a regular clock, as shown in **Figure 11**.

A 1-PPS signal carries clock synchronization information and is passed down from the spine or fabric switch to the ToR switch, and then to the NIC or SmartNIC. You may need a 1-PPS buffer or level translator in the active

cable paddle card for level translation and to generate additional copies.

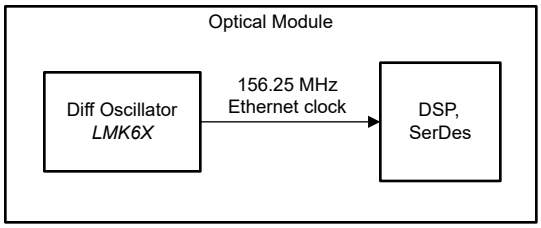

Figure 10. *Example optical module clock architecture.*

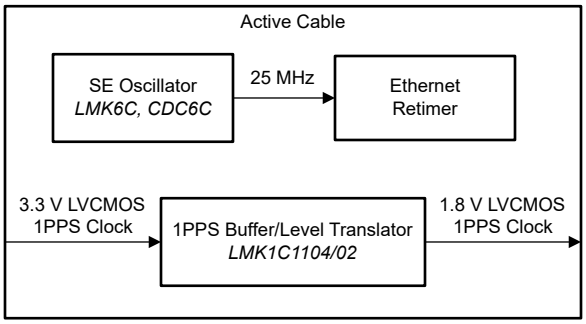

Figure 11. *Example active cable clock architecture.*

Generating the reference clocks for high-speed Ethernet SerDes requires an extremely low-jitter clock generator, as shown in **Figure 12**. A spine or fabric switch also requires similar or better Ethernet clock performance compared to a ToR switch. In addition, you'll need a timing digital phase-locked loop (DPLL) for network synchronization, as shown in **Figure 13**.

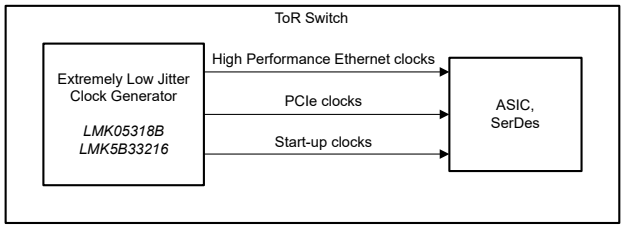

Figure 12. *Simplified ToR switch clock architecture.*

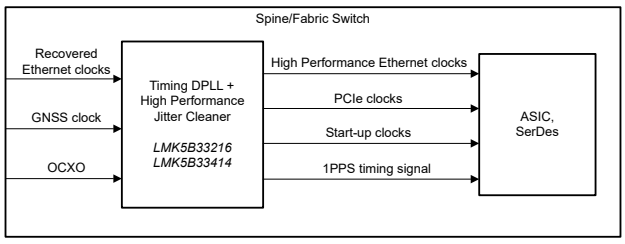

Figure 13. *Simplified spine or fabric switch clock architecture.*

Trend toward lower jitter

Lower jitter is necessary for high-speed Ethernet applications such as ToR switches, spine or fabric switches, optical modules, and NICs or SmartNICs. 112-Gbps-lane-speed Ethernet SerDes typically requires 125- or 100-fs 12-kHz to 20-MHz root-mean-square jitter at 156.25 MHz. 224-Gbps Ethernet typically requires 70 fs; future Ethernet clocks should achieve 50-fs maximum jitter or better. TI's proprietary bulk acoustic wave (BAW) technology provides 65-fs maximum jitter at 156.25 MHz in current TI products. Reference [1] provides more details about BAW technology.

Jitter requirements for PCIe reference clocks are also getting more stringent, especially in PCIe Gen 6, where the modulation scheme changed from non-return-to-zero (NRZ) to pulse amplitude modulation 4-level (PAM-4). Because PAM-4 operates on four levels compared to two levels in NRZ, it requires lower noise from the reference clock. This is also the reason why 56-Gbps PAM-4 Ethernet requires significantly better jitter performance than 28-Gbps NRZ. However, because PCIe compliance defines a noise transfer function to "filter" the jitter, the jitter requirement for PCIe is much more relaxed than Ethernet.

Greater integration

Another trend in data centers is clock integration, which enables greater reliability, smaller size and lower cost. Area reduction is especially important on daughtercards, where space is limited.

One important step when integrating is to eliminate the external crystal resonator (XTAL) or crystal oscillator (XO). Some vendors provide ICs with integrated XTAL, but this integration has some disadvantages. First of all, crystal integration typically requires a special land-grid-array (LGA) package that has substrate inside, instead of the much simpler, industry-preferred quad flat no-lead (QFN) package. LGA packages cost more and are bad for solder inspection. Besides, stacking the crystal on top of the base die increases the package height. For

example, a regular QFN package is only 0.9 mm high. With crystal integrated, the package becomes as tall as 1.7 mm, which can become a concern for casing.

Integrating BAW avoids all of these problems. BAW not only provides sub-65-fs RMS jitter, but it is very small and low cost as well. Putting the BAW die on top of the base die does not require an LGA package; a regular 0.8-mm or 0.9-mm QFN package height is sufficient. Therefore, the cost of BAW integration is much lower than crystal integration. BAW is also less sensitive to vibration and has better aging performance and a much lower failure rate compared to crystals. Data and details can be found in references [2] and [3].

Integration also enables designers to combine buffers, multiplexers and local clocks. A single integrated circuit can generate local PCIe clocks, buffer or multiplex external PCIe clocks, and level-translate 1-PPS signals at the same time. Instead of using multiple clock generators, multiplexers, buffers and oscillators, one clock generator can meet all of your clocking needs.

Conclusion

Data center clock trees are getting more complex and need a wide variety of clocking devices including oscillators, buffers, multiplexers, clock generators and network synchronizers. Some applications require low cost and low performance clocks, while others demand extremely high performance that few vendors can achieve.

TI aims to provide a complete clocking portfolio that meets any data center clocking requirement, simplifies clock trees through greater integration, improves system performance with lower jitter and reduces the bill of materials (BOM) cost leveraging the BAW technology.

References

1. Texas Instruments: ***TI BAW technology enables ultra-low jitter clocks for high-speed networks***
2. Texas Instruments: ***High Reliable BAW Oscillator MTBF and FIT Rate Calculations***
3. Texas Instruments: ***Standalone BAW Oscillators Advantages Over Quartz Oscillators***
4. **Clock buffers**
5. **Clock generators**
6. **Clock jitter cleaners & synchronizers**
7. **Oscillators**

IMPORTANT NOTICE AND DISCLAIMER

TI PROVIDES TECHNICAL AND RELIABILITY DATA (INCLUDING DATA SHEETS), DESIGN RESOURCES (INCLUDING REFERENCE DESIGNS), APPLICATION OR OTHER DESIGN ADVICE, WEB TOOLS, SAFETY INFORMATION, AND OTHER RESOURCES “AS IS” AND WITH ALL FAULTS, AND DISCLAIMS ALL WARRANTIES, EXPRESS AND IMPLIED, INCLUDING WITHOUT LIMITATION ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT OF THIRD PARTY INTELLECTUAL PROPERTY RIGHTS.

These resources are intended for skilled developers designing with TI products. You are solely responsible for (1) selecting the appropriate TI products for your application, (2) designing, validating and testing your application, and (3) ensuring your application meets applicable standards, and any other safety, security, regulatory or other requirements.

These resources are subject to change without notice. TI grants you permission to use these resources only for development of an application that uses the TI products described in the resource. Other reproduction and display of these resources is prohibited. No license is granted to any other TI intellectual property right or to any third party intellectual property right. TI disclaims responsibility for, and you will fully indemnify TI and its representatives against, any claims, damages, costs, losses, and liabilities arising out of your use of these resources.

TI’s products are provided subject to TI’s Terms of Sale or other applicable terms available either on ti.com or provided in conjunction with such TI products. TI’s provision of these resources does not expand or otherwise alter TI’s applicable warranties or warranty disclaimers for TI products.

TI objects to and rejects any additional or different terms you may have proposed.

Mailing Address: Texas Instruments, Post Office Box 655303, Dallas, Texas 75265