AI/SC Workstation Build Part 1.

Understanding AI/SC Workstation Architecture

Introduction:

This document will provide a comprehensive deep dive into the world of Commercial and Enterprise compute hardware; exploring the requirements to run modern AI and Scientific Compute workloads. This is part 1 of a document series that will enable the audience to scope out and construct a powerful AI workstation from a wide variety of hardware. This guide will allow developers and educators around the world to construct workstations at nearly any price point with superior compute and VRAM compared to the most expensive Commercial solutions.

AI/SC Workstation Build Catalog

SECTIONS

What are the limitations of current Commercial Systems?

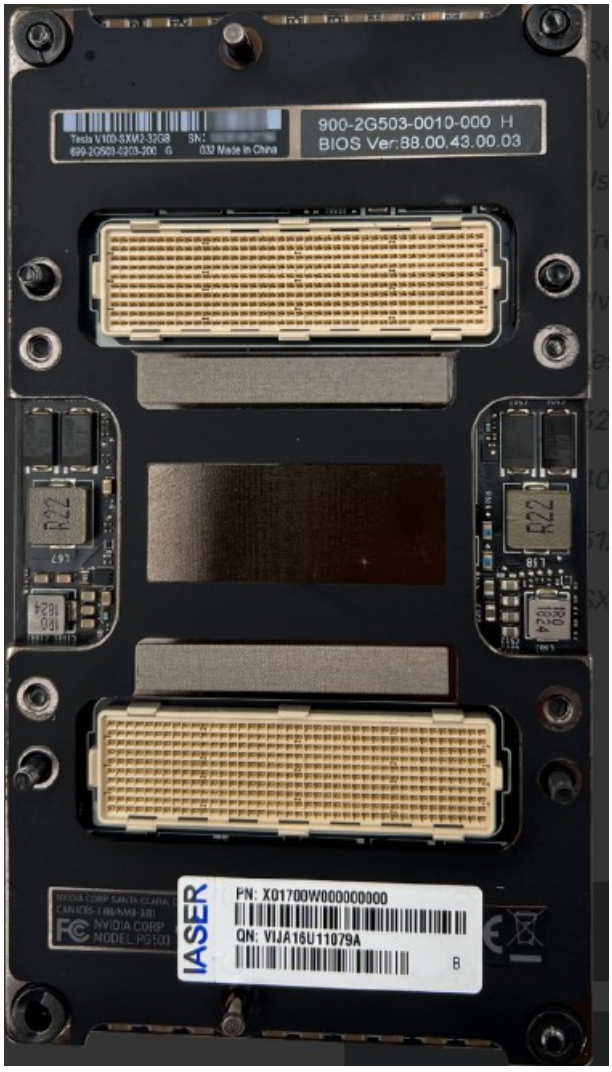
When discussing “Commercial” systems it is in the context of traditional PCIe GPU’s currently sold on the global market. Today in early 2026 a 96gb RTX Pro Blackwell 6000 GPU will set you back +$8,000 or a 4090 or even 5090 at a +$2,000 with only 32gb of VRAM. These can be entirely written off as candidates for FP64 Scientific Compute tasks since they both are outmatched by a Tesla K80 GPU from 2014. (**K-80 FP64: 1.87 Tflops** vs **Blackwell 6000’s FP64: 1.71Tflops with only 384 FP64 cores**.

FP64 aside and cards like the 5090 and Blackwell 6000 excel in most tasks like running medium to smaller sized LLM’s, Open Sora 2 etc. These GPU’s have excellent int8, TF16, TF32 performance overtaking GPU’s like the A100. Commercial solutions do 90% of what most GPU’s are tasked with doing for the general public. Only when we begin to scrutinize things like VRAM Bandwidth, Capacity, Scalability do older Nvidia Enterprise leaning architectures such as Pascal, Volta, Ampere, and the current Blackwell & Hopper are unmatched.

NVLink and AMD’s xGMI were the Enterprise successors to the phased out SLI and Crossfire data bridge technologies. NVLink 2.0 for example with the Volta generation can have a large 8x V100 32gb Topology and even beyond with the use of NVLink switches. Allowing for a cumulative 256gb of HBM2 memory with the following cumulative bandwidth. 900gb/s per V100 * 8 GPU’s = **7,200gb/s** with 300gb/s for GPU to GPU NVLink.

Later generations such as Ampere and Hopper took this to the next level with the use of NVLink switches to facilitate higher speeds and the ability to stack 8 trays of 8 H200’s for a total of **9,024gb** of HBM3.
All courtesy of NVLink and the SXM (Server PCI express module) socket standard. We will be levering some of these capabilities at home without spending $80k but only needing between $850 to $6,000 based on build type.

If you want to run massive AI models of any type, up to Llama 405b int-4 entirely on unified HBM2+ (+256gb) GPU VRAM or beyond with CPU offloading. This guide explores only the most economical solutions for a problem that limits the general public into the high performance computing space rivaling small business infrastructure.



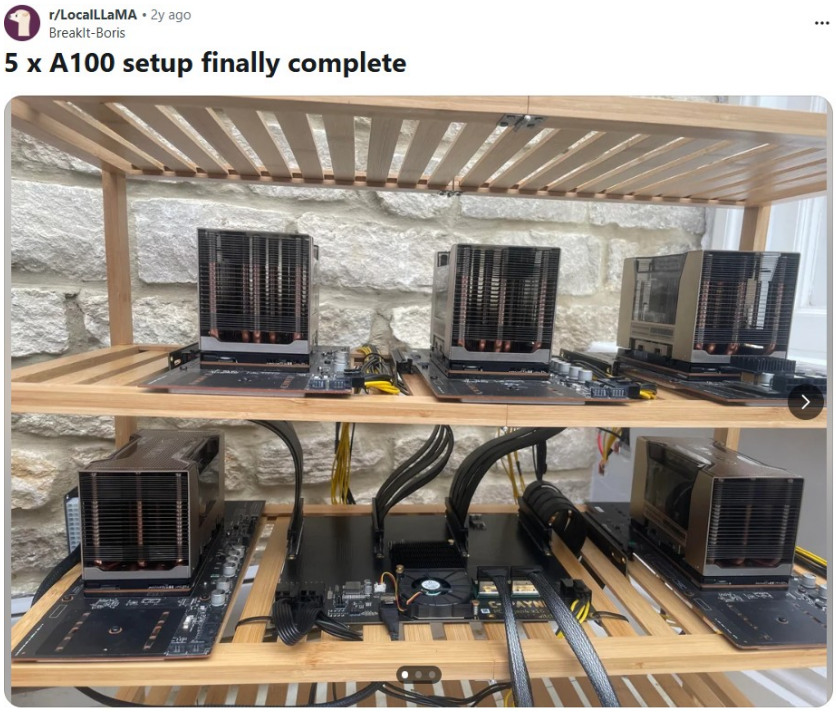
Underside of a SXM2 V100 GPU showing the dual 400 pin Mezzanine Amphenol FCI MEG-ARRAY

A Pair of H100 PCIe GPU's utilizing triple NVLink bridge connectors on the top.

Why has the PCIe GPU standard been replaced by Mezzanine based solutions?

Starting with the Tesla P100, Nvidia's SXM (Server PCI express module) was a revised physical standard that intended to optimize three domains. Thermal management, power delivery, and NVLink communication. NVLink has been utilized in the standard PCIe format and continues to be utilized for Ampere and Hopper though at slower NVLink speeds compared to GPU to GPU communication on the SXM4 and SXM5 standard. This trend has been followed by AMD with their AMD Instinct MI200 series of compute GPU. Mezzanine connected GPU's deliver more power to the GPU Die, provide superior data connectivity, and provide a more efficient cooling solution within the data-center setting.

Reddit user that converted 5 SXM4 A100 GPU's to a PCIe PEX based data truncation board. (No NVLink)



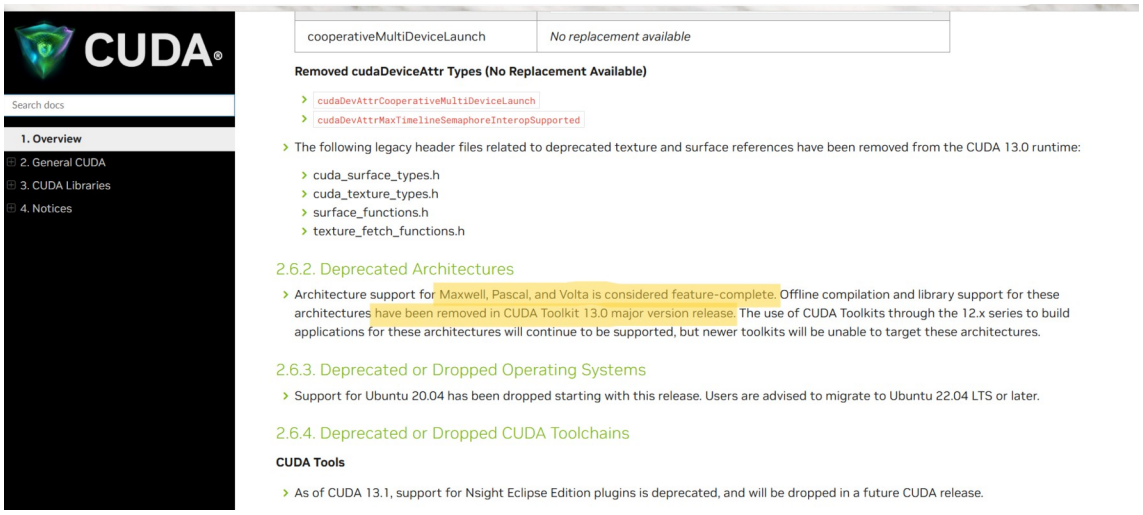
Why do we NEED the SXM standard?

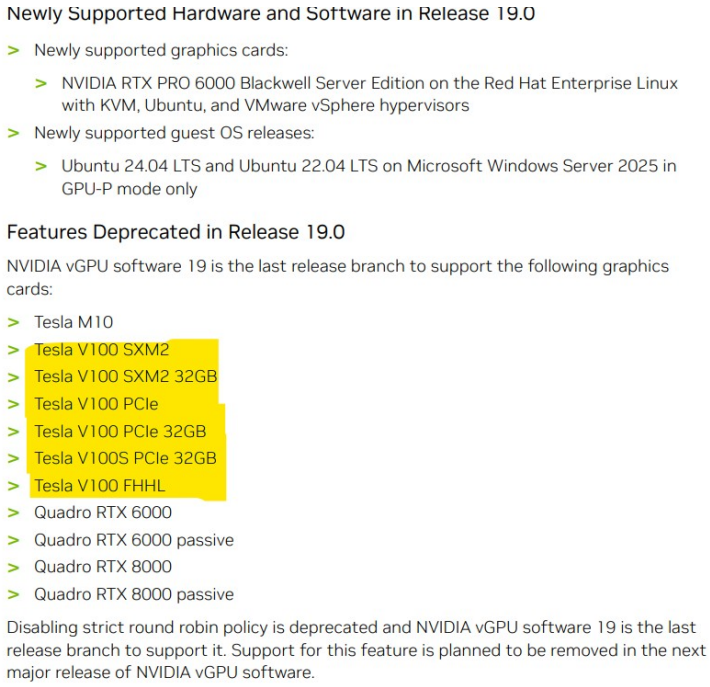
This obviously is based on your application, a converted V100 or A100 will be sufficient horsepower and VRAM for your requirements. If you want to host absolutely MASSIVE AI models, you will need to utilize traditional SXM host boards. Some notable examples of host board this project will work with are the following. Chinese Dual SXM2 Module, Super-micro's AOM-SXMV, and the HP P01786-b22 aka the XL270D SXM2 host board.
These boards are all Passive in nature despite some featuring PEX8749 PCIe truncation switches. Both the AOM-SXMV quad SXM2 board (2 PEX8749) the octa XL270D (4 PEX8749) has the same number of bi-directions PCIe3.0x16 outputs despite hosting 4x or 8x GPU's respectively. Even though in theory these PEX switches can in theory handle four V100s operating at PCIe3.0x16 in parallel, bandwidth may not reach full bandwidth. When hosting LLM's most of the communication relies on the unified NVLink 2.0 topology on these host boards. Each GPU provides six NVLinks for GPU to GPU communication to achieve a large unified Video Memory Pool. In addition to achieving a unified VRAM Pool, there are several other benefits SXM2 NVLink 2.0 provides. Each GPU's HBM2 memory compounds for a higher cumulative GPU bandwidth. Each of the six NVLink data lanes provided 50gb/s of bi-direction data transfer between GPU's. The P100 and V200 both are able to achieve NVLink without NVLink switches, future generations of Ampere, Hopper, Blackwell rely on NVLink switches to accommodate the massive Bandwidth increases.

Later variants of the SXM2 host boards utilized NVLink switches to grow beyond eight V100s, up to 16 GPU's (512 Gb of unified HBM2 if using 32Gb V100s). This was primarily optimized with the release of the SXM3 socket, featuring V100s with almost 100W more TDP than their SXM2 brothers.

You can have the fastest H200 on Earth with 141gb of 4.8Tb/s HBM3e Video Memory. An eight V100 SXM2 tray will have a unified 256Gb of HBM2 at nearly 7.2TB/s. The cost difference between the two options is $40k for the H200 and only $5000 for eight V100s and the SXM host tray.

Unfortunately Nvidia just deprecated the V100 drivers and CUDA Toolkit Releases with the R580 Driver package and CUDA Toolkit 13.0





The V100 gets an ACT II

The V100 was released in June of 2017, it is going to stay with us for a long long time. Currently around the world the V100 remains a large portion of the GPU compute data-center architecture. Currently the A100, H100/200, and the new Blackwell chips are incredibly expensive and are a Hot commodity for corporate acquisition teams. V100 has excellent computational potential, with eight V100s can match the Blackwell GB-100 in several computation parameters. These are very power efficient and compute dense GPU's, despite their age. People have been able to push these to their absolute limits using superior modern thermal pads and paste. I personally opt to use Arctic MX-4 due

to its exceptional stability and superior thermal conductivity to Graphite Pad technology used back in 2017. I find the TP-3 Pads from Arctic also do a great job of maintaining stable core and memory clocks in addition to a well cooled GPU Die.

Despite setbacks in most recent support in 2025 for CUDA Toolkit and the R580 Drivers, these remain an amazing cost effective option to do just about anything you want. Open Sora 2 with an additional up-scaling 4k model? YES! Hosting a massive LLM like Mixtral8x22B or Saily 220B models entirely on VRAM? YES!

As V100 prices continue to plummet like a rock, including their server host boards. The general public has access to tons of powerful Tensor compute on the resale market. Most Home Lab developers and educators opt to use Commercial GPU's or entirely host AI models entirely on RAM and use CPU compute instead. The **Commercial option** limits your VRAM and also your FP64 scientific compute is non-existent. The CPU only option provides Terra-bytes of potentially amazing RAM, but despite Intel's (VNNI, AMX, DL Boost, and the AVX-512) x86 instruction sets, it will still be astronomically slow compared to GPU based systems.

Thus the cockroach of data centers lives on to serve another generation of Developers and Educators with this guide!

The only limiting factor that stands in our way is the following.

- **SXM Enterprise trays are still gatekept, thus this option has not taken off yet at scale for home lab developers.**
- **Companies will sometimes Brand Lock trays to render them unusable.**
- **Create confusing standards like Exa-MAX for the latest Ampere and Hopper host trays for PCIe output and NVLink tray bridges. Or the Supermicro JPCIe connector standard that we recently discovered a PCIe adapter PCB Gerber file for! (Found here (https://github.com/SKARN-eng/SKARN-Open-Source-Project-/blob/main/Schematics%20and%20Adapters/AOM-SXMV%20PCIe%20Adapter.zip) with credit to the original developer provided and link to original Chinese post.)**
- **Relatively straight forward power solutions that are very easy to overcome. (This is only a negative due to lack of online documentation from manufacturers)**

We figure it out.

The **Chinese Dual SXM2 Trays for $250** This has been figured out already. Just connect two pairs or just two PCIe SFF-8654 to a (quad SFF-8654 8i PEX8749 PCIe adapter card) or a (dual SFF-8654 PCIe passthru card). They figured it out for you. Cheapest option on the entire market.

The **AOM-SXMV** for example is a quad SXM2 as mentioned previously. The JPCIe adapter essentially converts this back to standard PCIe3.0X16 acting just as a PCB passthru connector. The board has only two PCIe3.0x16 outputs so most Enterprise and enthusiast level Xeon or Threadripper/EPYC CPU's will work great. Even the ancient X99 platform for LGA2011 can work exceptionally well even in single CPU configuration since it has sufficient PCIe lanes. Also the original server allocated each pair of V100s through respective PEX8749 chips to two separate CPU's. Either the second cheapest or third cheapest option on the market.
By using a Xeon Scalable (any generation) or Threadripper/EPYC we can essentially host the entire tray on a single CPU. Entirely eliminating NUMA dual CPU penalties due to Intel's QPI/UPI or AMD's InfinityFabric data throughput. So we can theoretically squeeze significantly more performance out of these SXM2 trays with more modern processors.

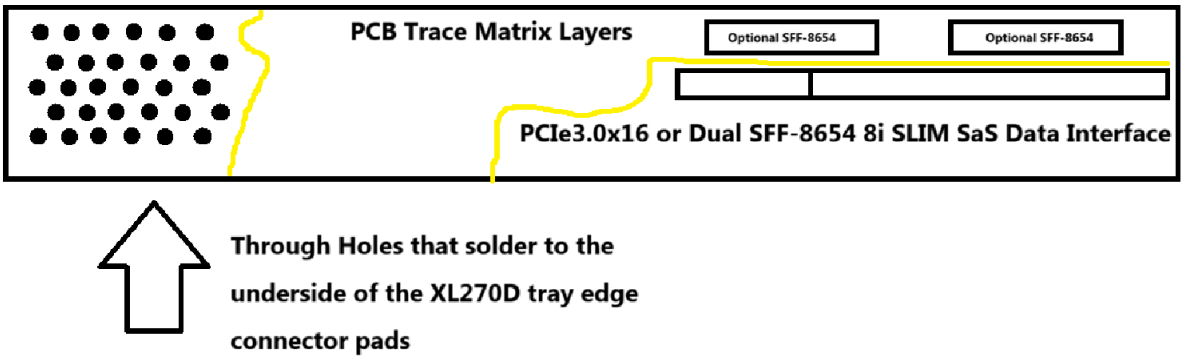
The **HPE XL270D** is a work in progress, at the time of writing this guide but you will find it here soon as it is finished!
(**https://github.com/SKARN-eng/SKARN-Open-Source-Project-/tree/main/Schematics%20and%20Adapters**)

1. Could be the cheapest and best deal ever. These trays are only $200 with some Ebay Enterprise sellers.

The goal there is to make a simple rectangular PCB design that simply provides large through-hole VIA's the can be placed to the underside of the edge connector pads and filled with solder. The board uses two 250 pin PCIe connections in conjunction with two pairs of PEX8749 switching IC's. (This board also has in addition to 8x SXM2 slots, it features a further **four** full PCIe3.0x16.) This PCB will just ac as a pass-thru connector like the original Mid-plane connector for this server. These trays can also be partially filled so you don't need the full eight V100s, you can build that number up with one per two paychecks. Starting with two V100s is a smaller cost burden than a full tray potentially costing $4000 total.

If someone beats me to this, I and the entire community owe a great deal of gratitude to you!

Here is my idea for the XL270D and also the ExaMAX Connectors on the A100 and H200 NVLink host boards. Please give this a try and let us all know how this works out.



This will be something that comes as a bare board or can be customized when ordering from sites like JLCPCB or MacroFab. Either with or without the X16 or SFF connector soldered on.

SXM4 and Beyond…..

When it comes to the ExaMAX connectors found on the Ampere and Hopper host boards. These boards usually output non-truncated PCIe4.0 and 5.0 x16 for each SXM GPU on the board. Thus we either will need to figure out the latest PCIe generation switching IC's or… we can just not solder all the pins. Thus we can have 8x H100's that offer 128PCIe lanes, and **only need to use 64 if we soldered just enough pins to allow for PCIe x8 for each GPU. Most of the communication happens between individual GPU's rather than as frequently as the CPU.**
Essentially sacrificing some PCIe lanes to have PCIe lanes for more M.2 storage and any additional PCIe cards….or more SXM trays.

This is still a theoretically straightforward solution to a real compute accessibility problem. This is a realistic and affordable option to get a depreciated 8x A100 80gb (640Gb Unified VRAM) fully loaded tray working with almost any motherboard if you have 54-year-old middle management money here in early 2026.

License CERN

1 Definitions

1.1 'Licence' means this CERN-OHL-P.

1.2 'Source' means information such as design materials or digital code which can be applied to Make or test a Product or to prepare a Product for use, Conveyance or sale, regardless of its medium or how it is expressed. It may include Notices.

1.3 'Covered Source' means Source that is explicitly made available under this Licence.

1.4 'Product' means any device, component, work or physical object, whether in finished or intermediate form, arising from the use, application or processing of Covered Source.

1.5 'Make' means to create or configure something, whether by manufacture, assembly, compiling, loading or applying Covered Source or another Product or otherwise.

1.6 'Notice' means copyright, acknowledgement and trademark notices, references to the location of any Notices, modification notices (subsection 3.3(b)) and all notices that refer to this Licence and to the disclaimer of warranties that are included in the Covered Source.

1.7 'Licensee' or 'You' means any person exercising rights under this Licence.

1.8 'Licensor' means a person who creates Source or modifies Covered Source and subsequently Conveys the resulting Covered Source under the terms and conditions of this Licence. A person may be a Licensee and a Licensor at the same time.

1.9 'Convey' means to communicate to the public or distribute.

2 Applicability

2.1 This Licence governs the use, copying, modification, Conveying of Covered Source and Products, and the Making of Products. By exercising any right granted under this Licence, You irrevocably accept these terms and conditions.

2.2 This Licence is granted by the Licensor directly to You, and shall apply worldwide and without limitation in time.

2.3 You shall not attempt to restrict by contract or otherwise the rights granted under this Licence to other Licensees.

2.4 This Licence is not intended to restrict fair use, fair dealing, or any other similar right.

3 Copying, Modifying and Conveying Covered Source

3.1 You may copy and Convey verbatim copies of Covered Source, in any medium, provided You retain all Notices.

3.2 You may modify Covered Source, other than Notices.

You may only delete Notices if they are no longer applicable to the corresponding Covered Source as modified by You and You may add additional Notices applicable to Your modifications.

3.3 You may Convey modified Covered Source (with the effect that You shall also become a Licensor) provided that You:

a) retain Notices as required in subsection 3.2; and

b) add a Notice to the modified Covered Source stating that You have modified it, with the date and brief description of how You have modified it.

3.4 You may Convey Covered Source or modified Covered Source under licence terms which differ from the terms of this Licence provided that You:

a) comply at all times with subsection 3.3; and

b) provide a copy of this Licence to anyone to whom You Convey Covered Source or modified Covered Source.

4 Making and Conveying Products

You may Make Products, and/or Convey them, provided that You ensure that the recipient of the Product has access to any Notices applicable to the Product.

5 DISCLAIMER AND LIABILITY

5.1 DISCLAIMER OF WARRANTY -- The Covered Source and any Products are provided 'as is' and any express or implied warranties, including, but not limited to, implied warranties of merchantability, of satisfactory quality, non-infringement of

third party rights, and fitness for a particular purpose or use are disclaimed in respect of any Source or Product to the maximum extent permitted by law. The Licensor makes no representation that any Source or Product does not or will not infringe any patent, copyright, trade secret or other proprietary right. The entire risk as to the use, quality, and performance of any Source or Product shall be with You and not the Licensor. This disclaimer of warranty is an essential part of this Licence and a condition for the grant of any rights granted under this Licence.

5.2 EXCLUSION AND LIMITATION OF LIABILITY -- The Licensor shall, to the maximum extent permitted by law, have no liability for direct, indirect, special, incidental, consequential, exemplary, punitive or other damages of any character including, without limitation, procurement of substitute goods or services, loss of use, data or profits, or business interruption, however caused and on any theory of contract, warranty, tort (including negligence), product liability or otherwise, arising in any way in relation to the Covered Source, modified Covered Source and/or the Making or Conveyance of a Product, even if advised of the possibility of such damages, and You shall hold the Licensor(s) free and harmless from any liability, costs, damages, fees and expenses, including claims by third parties, in relation to such use.

6 Patents

6.1 Subject to the terms and conditions of this Licence, each Licensor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable (except as stated in this section 6, or where terminated by the Licensor for cause) patent licence to Make, have Made, use, offer to sell, sell, import, and otherwise transfer the Covered Source and Products, where such licence applies only to those patent claims licensable by such Licensor that are necessarily infringed by exercising rights under the Covered Source as Conveyed by that Licensor.

6.2 If You institute patent litigation against any entity (including a cross-claim or counterclaim in a lawsuit) alleging that the Covered Source or a Product constitutes direct or contributory patent infringement, or You seek any declaration that a patent licensed to You under this Licence is invalid or unenforceable then any rights granted to You under this Licence shall terminate as of the date such process is initiated.

7 General

7.1 If any provisions of this Licence are or subsequently become invalid or unenforceable for any reason, the remaining provisions shall remain effective.

7.2 You shall not use any of the name (including acronyms and abbreviations), image, or logo by which the Licensor or CERN is known, except where needed to comply with section 3, or where the use is otherwise allowed by law. Any such permitted use shall be factual and shall not be made so as to suggest any kind of endorsement or implication of involvement by the Licensor or its personnel.

7.3 CERN may publish updated versions and variants of this Licence which it considers to be in the spirit of this version, but may differ in detail to address new problems or concerns. New versions will be published with a unique version number and a variant identifier specifying the variant. If the Licensor has specified that a given variant applies to the Covered Source without specifying a version, You may treat that Covered Source as being released under any version of the CERN-OHL with that variant. If no variant is specified, the Covered Source shall be treated as being released under CERN-OHL-S. The Licensor may also specify that the Covered Source is subject to a specific version of the CERN-OHL or any later version in which case You may apply this or any later version of CERN-OHL with the same variant identifier published by CERN.

7.4 This Licence shall not be enforceable except by a Licensor acting as such, and third party beneficiary rights are specifically excluded.