OPEN

Compute Project

Proposed Charter For Virtual IO

Revision History

Date	Name	Description
01/19/2012	James Hesketh Joel Winland DaWane Wanek	Original draft based upon Template from Systems Management Track project. Various edits to bring up to useable draft.
02/27/20 12	James Hesketh	Incorporate Feedback and decisions from the Conference call on the charter - 02-16-2012
08/14/2012	James Hesketh	Revise Charter for Incubation Committee review.
08/24/2012	DaWane Wanek, Carl Mies, Steve Knodl, Stephen Rousset, Joel Wineland ,James Hesketh	Discussion and changes to the charter to present to the incubation committee.
09/01/2012	DaWane Wanek, Carl Mies, Steve Knodl ,Joel Wineland ,James Hesketh, Steve Sommers	Discussion and changes to the charter to present to the incubation committee.

Summary

The Open Compute project has a mandate to provide scalable compute resources and infrastructure in an Open manner aligning with Open Source tenants.

In a scale environment there is a tipping point where the demands of "scaling" introduce multifaceted problems with supply chain activities that require substantial time and investment to manage. Operational costs and component life-cycles have posed barriers to operational efficiency and capital costs; the approach of tailoring hardware designs to reduce complexity and maximize efficiency is a viable means of offsetting some of this cost.

The Virtual IO group believes that being able to combine and re-combine resources such as compute, memory, networking and storage (to name the most pressing) to fit a required need offers a new path to obtain greater efficiency. In short Virtual IO has the ability to reduce the support burden, reduce the time to deliver a usable IT framework, and reduce the financial overhead of operating at scale; and do so utilizing open standards that are consistent in approach and application.

License

As of April 7, 2011, the following persons or entities have made this Specification available under the Open Web Foundation Final Specification Agreement (OWFa 1.0), which is available at http://www.openwebfoundation.org/legal/the-owf-1-0-agreements/owfa-1-0 :

Facebook, Inc.

You can review the signed copies of the Open Web Foundation Agreement Version 1.0 for this Specification at http://opencompute.org/licensing/, which may also include additional parties to those listed above.

Your use of this Specification may be subject to other third party rights. THIS SPECIFICATION IS PROVIDED "AS IS." The contributors expressly disclaim any warranties (express, implied, or otherwise), including implied warranties of merchantability, non-infringement, fitness for a particular purpose, or title, related to the Specification. The entire risk as to implementing or otherwise using the Specification is assumed by the Specification implementer and user. IN NO EVENT WILL ANY PARTY BE LIABLE TO ANY OTHER PARTY FOR LOST PROFITS OR ANY FORM OF INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES OF ANY CHARACTER FROM ANY CAUSES OF ACTION OF ANY KIND WITH RESPECT TO THIS SPECIFICATION OR ITS GOVERNING AGREEMENT, WHETHER BASED ON BREACH OF CONTRACT, TORT (INCLUDING NEGLIGENCE), OR OTHERWISE, AND WHETHER OR NOT THE OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Overview

The goal of this Project is to provide devices capable of Virtual IO (VIO) within an Open Compute framework.

Additionally specific methodologies, technologies or frameworks of VIO will be defined in work-streams within the VIO project.

Through previous design and feedback sessions two work streams have already been categorized, these are:

Exclusive IO Work stream

The categorization of “Exclusive IO” is that IO data paths are abstracted from the root complex (enclosure) and moved outside to an external device. This will be in a one (root complex) to one or many devices.

Exclusive IO Work-stream

The primary defining criterion of this category is that a normally closely coupled resource(s) bound to a root complex (such as storage HBA, or Network Card) is decoupled (externalized), but still able to maintain full functionality.

The secondary defining criterion of this category is that resources that are decoupled maintain a one to one relationship between root complex and IO resource.

The use of device virtualization methodologies or technologies is not a requirement in this category, as any device can be extended out of a chassis by physical means.

Please note that a SR-IOV or MR-IOV capable resource will fall into this category if deployed in such a manner as a one to one relationship with a VH is maintained. (See PCIe sig for definition of physical (VH).

Example.

A hyper-visor has three VM's, they are VM1, VM2 and VM3. The Network card is SR-IOV capable and has 5 virtual functions and 1 physical function. (See PCIe sig for definition of physical (PF) or virtual (VF) functions.)

The NIC presents the PF to the hyper-visor 1 and the hyper-visor passes 3 VF’s to each VM respectively. Any communication over a network is handled by the upstream access switch or the software switch within the hyper-visor. Therefore for VM1 to send a TCP packet to VM2, the packet is constructed on VM1 and transmitted to the software switch, the software switch will then forward the packet to VM2 where it is de-constructed.

It is worth noting that VF1 cannot bypass the software switch and send a packet directly to VF2. Furthermore as the SR-IOV capable NIC is bound to the hyper-visor VH it maintains a one to one relationship.

Shared IO Work stream

The categorization of “Shared IO” is that a resource can share its IO in a many to one arrangement.

The primary defining criterion of this category is that a single resource provides IO to multiple root complex(s) in a discrete manner.

The secondary defining criterion is that the resource must be presented in a transparent manner between end points.

The Third defining criterion is that in shared IO, connections should be made as if point to point. This is inferred as many to one.

The following illustrates these criteria in an example.

A Virtual IO device has three root complexes homed on it, they are Phys1, Phys2 and Phys3. The Network card is MR-IOV capable and has 5 physical functions. (See PCIe sig for definition of physical (PF).)

The NIC presents three PF’s to Phys1, Phys2 and Phys3 respectively. Any communication over a network from any of the Phys servers is done utilizing the PF's. However each PF will pass any traffic upstream to the access switch. r. Therefore for Phys1 to send a TCP packet to Phys2, the packet is constructed on Phys1's PF and transmitted to the access switch, the switch will then forward the packet to PF2 associated with Phys2 where it is de-constructed.

It is worth noting that PF1 cannot bypass the access switch and send a packet directly to PF2.

Goal

It is the intent of this VIO group to facilitate the delivery of both Extended IO and Shared IO to Open Compute, in this way we can cover multiple use cases with “building blocks” at scale. However the end goal and the aim is to provide combinations of Extended IO and Shared IO for multiple use cases, where it can be said an Open Compute VIO device can perform both Extended and Shared IO, or a combination of either.

VIO can have a multitude of use cases, most often in cost reduction.

However three compelling use cases present themselves.

1) Amortization. - If we can move closely coupled resources or move integrated components from within server chassis to outside the server chassis, then from a financial perspective we are able to amortize the cost of any given component over its respective given life-cycle.

   An example here is that storage (sub) systems have typically a longer life-cycle than a CPU. You can depreciate these components on different schedules.

2) Upgrade - If we can move closely coupled resources or move integrated components from within server chassis to outside the server chassis, then upgrade of components can be simplified, and existing components can be re-used. Today to upgrade a server chassis (from one model to a later model requires that the chassis be removed). In the VIO goal, this allows only the upgrade of the component that you wish independently of other components.

3) Logical Hardware configuration – If a robust set of API's are provided then the opportunity of configuring a specific hardware profile such as the amount of Network cards/ports, the type of Storage and the raw capacity (along with more long term aims of memory density and remote management cards,) is within reach. This layer of logic would reduce time to reconfigure and time to deploy within a scale environment. It may also assist with hardware failure remediation. This use case is a must to achieve a light or no touch data center.

Whilst the three use cases are all different, the approach and the method in developing hardware to answer the use cases questions are very similar. What will be in question is the technology/mechanism(s) used in the answers. Today PCIe is a mature standard that has provisions added to the standard to assist with the VIO endeavors; however it is not the only technology that is capable of delivering the required results.

For this reason we would like sub projects within the VIO working group. These sub projects with the question of technology and mechanism(s), where viable technologies will gain traction by real world implementations and graduation to a specification as per the Open Compute rules.

Work Streams of Virtual IO

These should be reviewed by the work stream leads and project chair. It is the responsibility of each project lead to represent the collective consensus and view of each sub-project's community and therefore output.

Focus Areas

There 4 focus areas this project needs to consider:

1) Fostering a community where technologies can be shared and explored:

It is conceivable that a method and manner of providing VIO will not change, however it is very unlikely this will hold true, as new technologies are explored and brought to market. A philosophy of not only *what are we doing now*, but *what do we want to do in the future and what should be used in the future* should be adopted. This philosophy translates to a work stream(s) around established technologies or meritorious concepts/ideas:

When an idea or concept gains merit within the community, we must be able to achieve the VIO project goals, as well as being able to bring this into a work stream of relevance.

2) Maturing an ecosystem around Resource(s):

It is not enough to be able to provide a framework or method of performing VIO with a device/infrastructure. A VIO device may give the functionality but a resource is provided and an End-point consumed. To make the VIO project a success, efforts must be put into providing an eco-system around resource and end point design and that eco-system should have an interest in on-going production of the solution. In other words, there should be an expectation of a broad appeal, at least within the Open Compute community. It is not sufficient to assume the maxim of "build it and they will come."

3) Identify execution partners for Projects or work streams :

For any output of a work stream to achieve an incubation status and move to a specification, an execution/manufacturing partner must be found. This is a simple tenant and is necessary so that the project at each phase is set up to be successful.

Each work stream will be responsible for considering the "how, why and who". The

"Why" should something be included in a work stream; "Who" is going to work in the work stream; and "How" are you going to design the output of the work stream. The Work stream should have a lead. Sufficient thought should be given to the fact that a project lead will viewed as the individual who may take a contributions in a work stream from embryonic, to incubation, then (hopefully) to specification. It is expected that this effort will be shared with the project chair.

Work streams should be distinct in the scope so that they are identifiable as a work-stream with merit and distinct value.

Meeting Cadence

To keep the VIO project and associated communities informed, we will adopt a bi-weekly meeting schedule. This meeting may be in-person or via other means such as teleconferencing whichever is agreed upon in advance. It is favorable that when summits or higher level Open Compute meetings take place, face-to-face meetings will be arranged.

Design and feedback sessions should be called to note and not included in the bi-weekly meetings, as these should be separate. A design or feedback session can be called for within any of the work streams, however to minimize disruption, efforts must be made to include sister projects (where applicable).

In all meetings, minutes and notes will be recorded please refer to the Open Compute by-laws on meeting minutes and how to publish them.

Organization

The Project will have roles; participation in any role requires the CLA be signed. The expected function of a role is outlined below.

1) Project Chairs

   The project chair will facilitate the flow of information, determine consensus and commit documents, as well as oversee (in coordination with the project leads) all sub-projects related to the VIO project.

2) Work stream Lead

   The project lead will facilitate the flow of information, determine consensus and commit documents, as well as oversee (in coordination with commit/working group members) any sub-project(s) related to the VIO project.

3) Commit/Working Group Members

   Commit/Working Group members are members of a sub-project that are working towards moving the project forward. Specifically, these members are tasked with moving agreed objectives forward between meetings. Examples of “working towards” can be advice or donation/contribution of Intellectual Property that pertains to the project or final specification. Donation and contribution of Intellectual Property can be achieved by either donating Intellectual Property that is legally belonging to the contributor under the licensing structure as referenced in the “Licensing” section, or by working within the working group members, where such work is already covered by the licensing structure as referenced in the “Licensing” section.

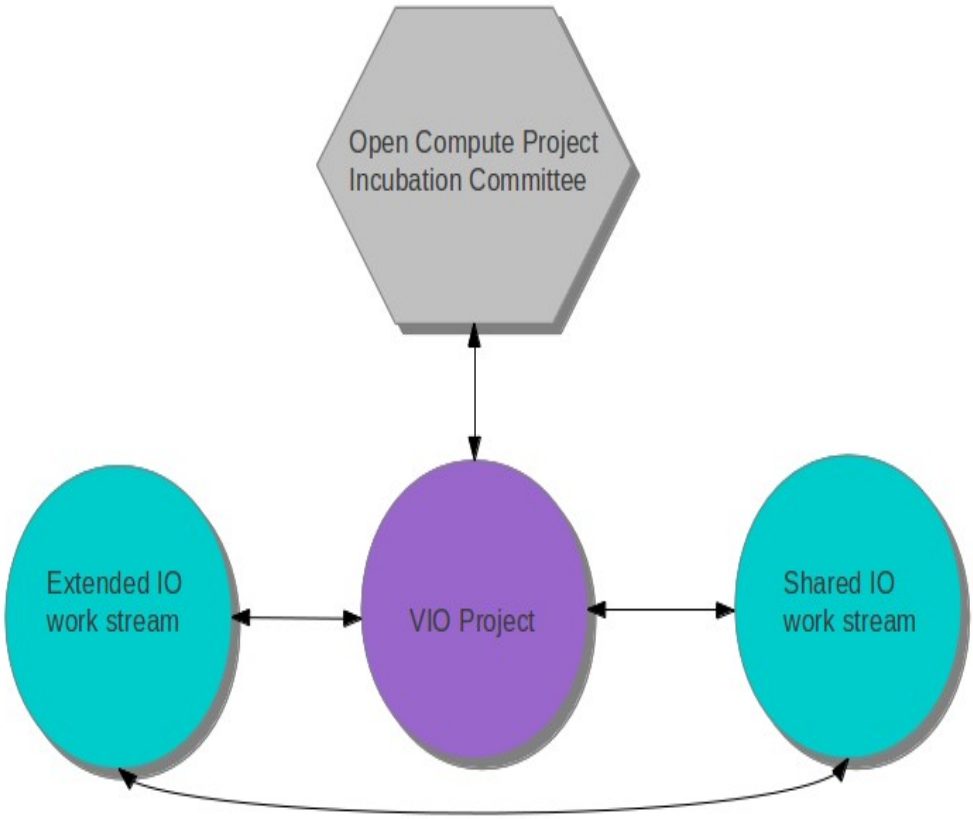
4) Advisory Members

   Advisory members are individuals who are providing input into a sub-project(s) and engaged with the meeting cadence defined within the sub-project.

As previously stated, some of the roles above are already defined within the Open Compute Foundation, however some are not. Specifically, the role of Working Group and Advisory are not and fall under a wider distinction. Assumption of these roles are generally self-defining, however any entry into any formal role will be at the Project lead’s discretion. Abrogation or removal from a role will require either the individual stepping down or a majority vote from the VIO sub-project community.

Additionally, there will be some individuals that will automatically become enrolled in such a role and cannot be removed by majority vote. These are limited to Open Compute committee members (either board or incubation) and Project chairs.

Organizational Chart



Appendix A

Definitions

Virtual IO System:

A system which provides either Exclusive IO or Shared IO functionality to IO Adapters.

IO Adapter(s):

An adapter which complies with the PCI Express standards. This is the component which will be virtualized.

PCIe Sig Definitions can be found here:

http://www.pcisig.com/specifications