OFA TAC: Technology Advisory Council Updates & Direction

Tom Stachura (Intel) & Deigo Crupnicoff (Mellanox)
April 24, 2013
TAC Charter

• OFA Charter:
  – Develop, test, license and distribute Open Fabrics Enterprise Edition (OFED) to deliver RMDA, kernel bypass and low latency fabric technologies
  – Promote industry awareness and acceptance of the above capabilities for a robust ecosystem of development and delivery

• TAC Charter
  – Investigate technology trends that could provide opportunities or roadblocks to the adoption of OFED technology
  – Review needs of end user markets/applications/new technology transitions focusing on impacts and opportunities for OFED
  – Maintain close relationships to IBTA TWG and other specification bodies, as well as end users
## OFA TAC/ IBTA TWG Interaction

### Key Focus
- Software Delivery
- SW/HW Alignment
- IB/Enet Specs

### Scope
- Markets/Applications
- Full Solution
- Systems/Nets

### Requirements
- End-user Needs
- Solution Architecture
- Technology Capability

### Enabling
- OS Distributions
- APIs (e.g. OFED)
- H/W Specifications

---

**Unchanged from 2012 Monterey Presentation**
1. Technology Review GASPI (PGAS API)
2. Embarked on investigation of OFA relevance
   – Starting with ULP dependencies/opportunities
3. Identified Key TAC Focus Areas
   3A) Cloud: OpenStack & OFED opportunities?
   3B) HPC: Verbs is “too heavy-weight”
1) GASPI

- TAC was asked to investigate GASPI:
  - GASPI is an open source PGAS API
  - Open source version of Fraunhofer’s GPI
    - http://www.gpi-site.com
- Fraunhofer ask of OFA:
  - Host GASPI to enable broader PGAS adoption
- Discussion:
  - OFA cannot favor one implementation over another
  - However, there is opportunity for win-win here…
Key Learnings from GASPI ask

• PGAS is an evolving OFA use model
  – It benefits OFA to start active engagement here

• OFA has opportunity to host ULP source
  – Does (& should not) imply inclusion in OFED package
  – However, it brings ULPs and OFED closer together
  – Thus, increasing OFED value and end-user experience

• ULP support also drives additional need and opportunity for stronger interoperability

Direct engagement w/ ULPs provides OFA opportunities
2) Seeking OFA Relevance

• OFA TAC is looking for ways to increase OFA relevance providing:
  – Better end-user experience
  – Additional membership
  – Increased OFED use

• GASPI experience clued the team into the value of the ULPs
  – TAC decided to explore the ULPs…

Need a deeper look at OFED-relevant ULPs
OFS is built on top of RDMA. (Not exclusively, but pretty much).

Applications are either coded to the Verbs API, or they rely on a ULP.

So evolving OFS may also mean evolving the network infrastructure that underlies it.

In other words, this isn’t solely an OFA problem.
Legacy, Data Analysis, Data Storage, & Data Access

Application layer

- IP apps
- skts apps
- Structured data Apps (DBs)
- Unstructured data apps
- Network filesystems (client/server)
- Block storage

Interface layer

- IPoIB
- SMC-r
- rsockets
- SDP*
- RDS
- uDaPL
- cluster file system
- NFS-RDMA
- SRP, iSER

Provider layer

- Kernel verbs consumer
- User verbs consumer

Hardware layer
Distributed Computing

Application layer
- Message Passing
- Shared Memory

Middleware
- GASNet Extended API

Interface layer
- MPI(s)
- Portals
- GNI
- Portals
- MPI(s)

Provider layer
- Kernel verbs consumer
- User verbs consumer

Hardware layer

Layers:
- Application layer
- Middleware
- Interface layer
- Provider layer
- Hardware layer

Providers:
- PSM
- uDaPL
- rsockets
- Other

APIs:
- MPI(s)
- Portals
- GASNet Extended API
3) Additional TAC Focus

- Worked to identify the top focus topics per market segment:
  
  **3A) HPC/Exascale:** Verbs is “too heavy-weight”
  - Invited an expert from Los Alamos (thanks Susan again!)
  - Output is basis of several of the workshop agenda items

  **3B) Cloud/EDC:** OFED & OpenStack synergy?
  - Invited experts from Mirtantis to TAC (thanks to Susan!)
  - Identified opportunities to engage
3A) HPC: “Heavy-weight” Verbs

• Nathan @ LANL shared his UD setup experience:
  1. RDMA-CM doesn’t scale
     – Could not scale to 1500 ranks; Issues seen @ 32 procs/node.
  2. RC mode runs out of queue pair resources
     – As discussed in the Dynamic Connection topic
  3. Verbs interfaces don’t map well to MPI semantics
     – As discussed in the PSM topic
  4. Verbs is heavyweight – a lot of coding & setup
     – e.g. Managing memory registration
  5. Lack of standardization between h/w implementations
     – e.g. PSM vs. MXM
  6. No “Well-known” ports
     – End-user desire to open specific ports, but QP # is random
3B) Cloud: OpenStack & OFED

• David & Jason @ Mirantis discussed OpenStack:
  – OpenStack: open source Cloud OS that defines APIs to control compute, storage and networking resources
    • Networking is TCP/IP
    • Storage is SCSI or TCP/IP
  – Guidance from Mirantis is to work to integrate/add OFED performance capabilities to targeted areas:
    1. Image migration (a.k.a. LANCE)
    2. iSCSI over RDMA (a.k.a. CINDER)
    3. Object Storage (a.k.a SWIFT)
    4. SR-IOV enabling
Next Steps & Feedback

• Next Steps
  – ULPs: Working to output a strategy & updated diagram
  – HPC: Identify actions given Workshop feedback
  – Cloud: Continue to investigate OpenStack opportunities and look for member contributions

• Feedback?
  – Other areas should the TAC be investigating?
    • Technologies to explore?
    • Experts to engage?
  – Additional desired outcomes/output?
BACKUP
TAC Focus Mindset

• Per Market
  – Trends:
    • End-use (e.g. Model/Simulation Needs)
    • Performance (e.g. 1 ExaFLOP)
    • Applications (e.g. Hadoop)
    • Hardware evolution (e.g. PCIe 4.0, Memory Bus)
  – Disruptors/Opportunities:
    • Alternate protocols (e.g. SHMEM)
    • New technologies (e.g. NVM)
    • New standards (e.g. OpenFlow)
    • New usages (e.g. FSI UDP/TCP Verbs)
  – Hear from industry experts
    • Invite experts inside AND OUTSIDE THE OFA