RDMAmojo: insights and thoughts of a blogger

OpenFabrics Software User Group Workshop

#OFSUserGroup

About me

- A Senior staff Engineer at Mellanox Technologies. I have more than 12 years experience in various roles such as:
  - a manager
  - a developer

- Wrote tens of applications over verbs
  - Over several verbs generations
  - In different OS's

- Was involved in several documentation projects in verbs programming
  - Man pages of libibverbs
  - 1st version of the "RDMA Aware Networks Programming User Manual"
  - Wrote the chapter "InfiniBand" in the "Linux Kernel Networking – Implementation and Theory" book by Rami Rosen, 2013 – Apress .

- Author of "RDMAmojo" - a blog on the RDMA technology

Problem

- ## There isn't much available information on RDMA
  - ### Theoretical information
    - There are specs, but same spec for provider and user
  - ### Practical information
    - Most of RDMA code is open source, but who wants to read the code?
    - The code itself doesn't always provide the usage

- ## If everything is working – great
  - But how to debug it without information?

- ## Available knowledge will increase adaptive of RDMA technologies

Solution: RDMAmojo blog

- Celebrating its 3rd anniversary
- Currently, I'm the only contributor
- 84 posts (>80 are technical)
  - RDMA packages deployment
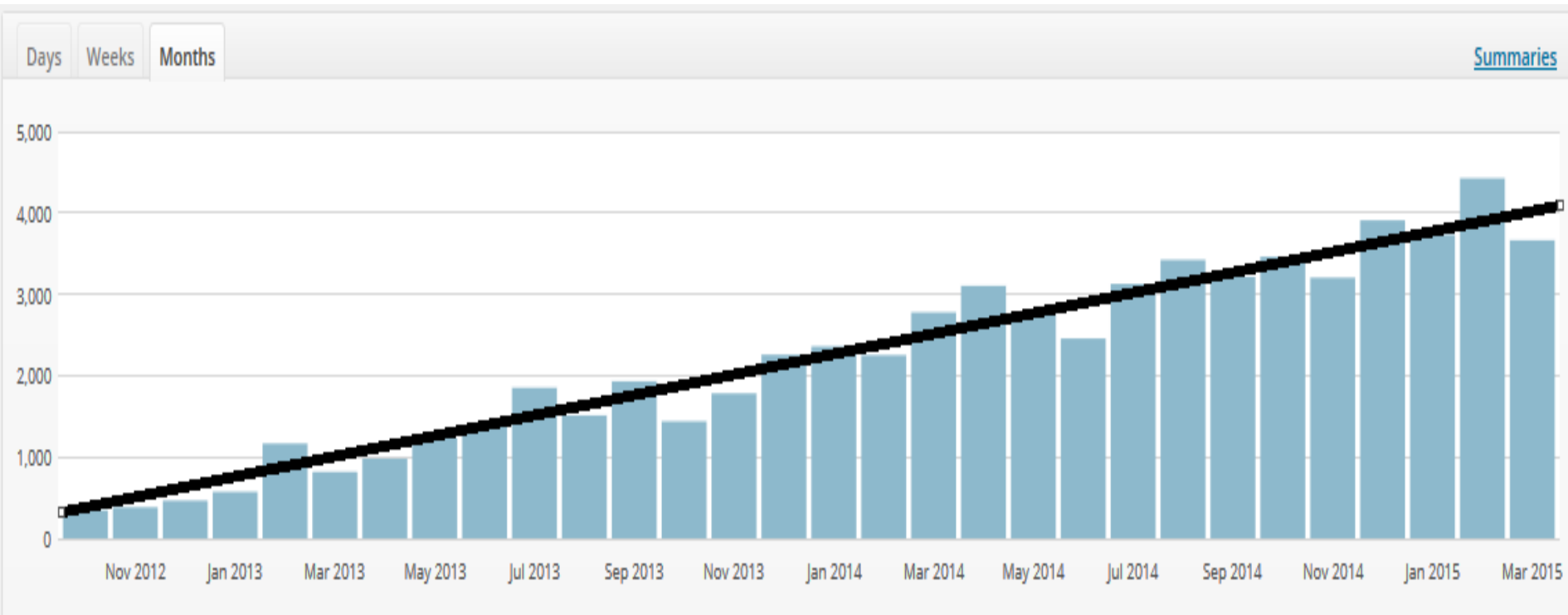 - RDMA fundamentals
  - libibverbs API

RDMAmojo // RDMAmojo – blog on RDMA technology and programming by Dotan Barak

RDMAmojo posts

- Cover an area
  - Verb
  - Feature
  - Flow
- Cover the relevant API
- Cover return values
  - Describe possible reasons for errors
- Give examples (when applicable)
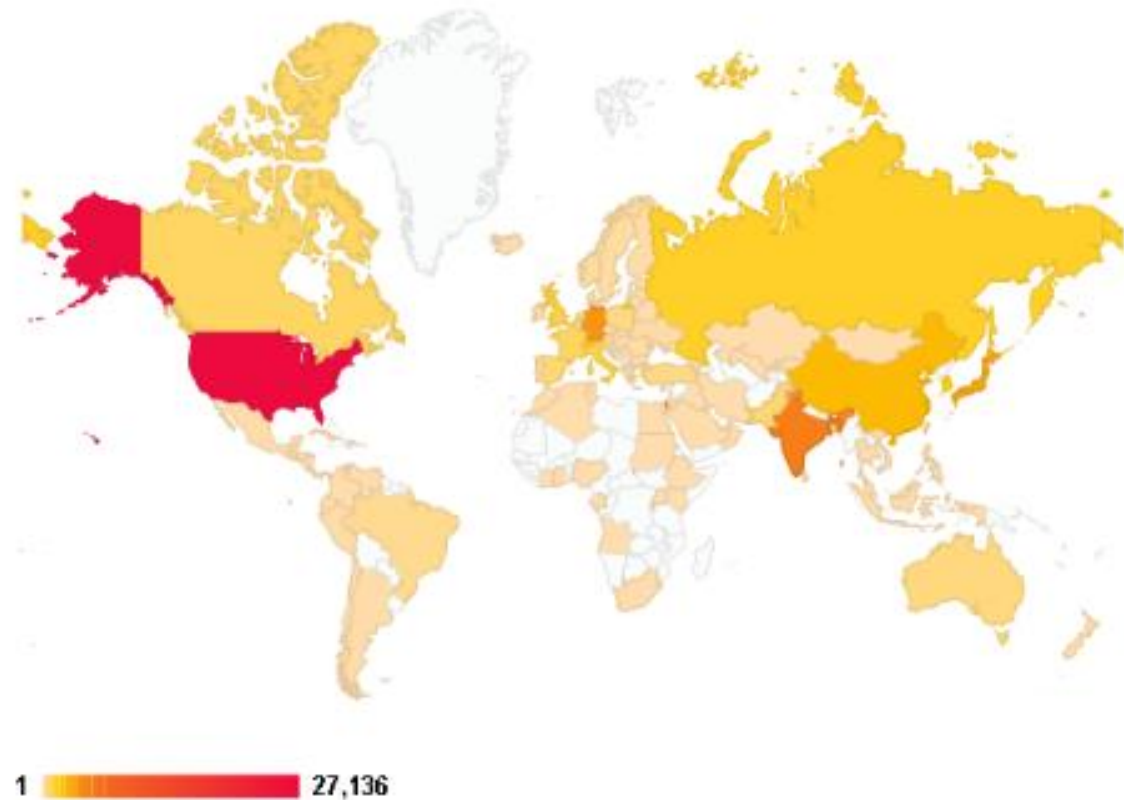- Q&A to interesting questions

- People can, and are, commenting on posts

views through time

views according to country

1. United states 27K
2. Israel 6.5K
3. India 5K
4. Germany 3.5K
5. Japan 3K

1   27,136

views per posts

1. ibv_post_send() 4.3K
2. ibv_poll_cq() 3.5K
3. Introduction to RDMA 2.5K
4. ibv_reg_mr() 2K
5. libibverbs 1.85K
6. ibv_modify_qp() 1.8K

ibv_post_send()

- There are MANY attributes
  - Depend on QP attributes
  - Depend on opcode

- There are tips and tricks
  - Performance improvements

- The feedback not always immediate

ibv_poll_cq()

- What is the meaning of all attributes?
- Sometimes there isn't any completion at all
- The meaning of status != success
- Not all attributes are always valid
  - Depend on the status
  - Depend on opcode

Introduction to RDMA/libibverbs

- There is a buzz on RDMA
- People wants to know what is it good for
- People wants a starting point for the technology
- Covers the libibverbs API
  - Role of every verb

ibv_reg_mr()

- Which memory can be registered?

- What to do when registration fails?
  - Debug
  - Increase resources

ibv_modify_qp()

- There are MANY attributes
    - Depend on QP attributes
    - Depend on transition
    - Depend on protocol
    - Some attributes are optional
- How to debug it?
    - EINVAL is always returned …

Frequency asked topics

- Connections establishment
- Data path flows
- Atomic

- Almost all questions were on InfiniBand/RoCE

Insights on RDMA users

- They are very smart
- In most cases they know exactly what they want
  - Just need some guidance how to do it
- They are doing very interesting things
- They adapt to changes quickly
  - New features
  - New operating systems

Future plans

- Cover more packages API
  - librdmacm
  - libibumad
- Cover management/tools
  - OpenSM
  - Diag tools
- Cover ULPs (protocol and usage)
  - IPoIB
  - RDS
- Cover RDMA related packages

What would you like to see?

Thank You

OpenFabrics Software
User Group Workshop