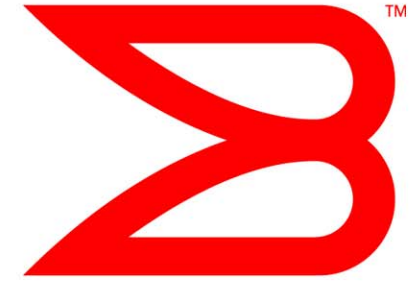


**BROCADE**



# **Data Center RDMA Protocol Study Group Proposal**

**03-18-09**

**09-141v0**

**Scott Kipp**

**Steve Wilson**

# Goals of Data Center RDMA Protocol

- Enable RDMA over CEE and FC physical layers without TCP
  - RDMA is used in high performance computing clusters to exchange memory between multiple computers without involving the operating system
  - TCP is designed for lossy-networks and carries significant overhead and latency on exchanges
  - Converged Enhanced Ethernet (CEE) and Fibre Channel (FC) provide very low loss networks and don't need TCP
- TCP does not scale well with low latency for RDMA exchanges over thousands of ports



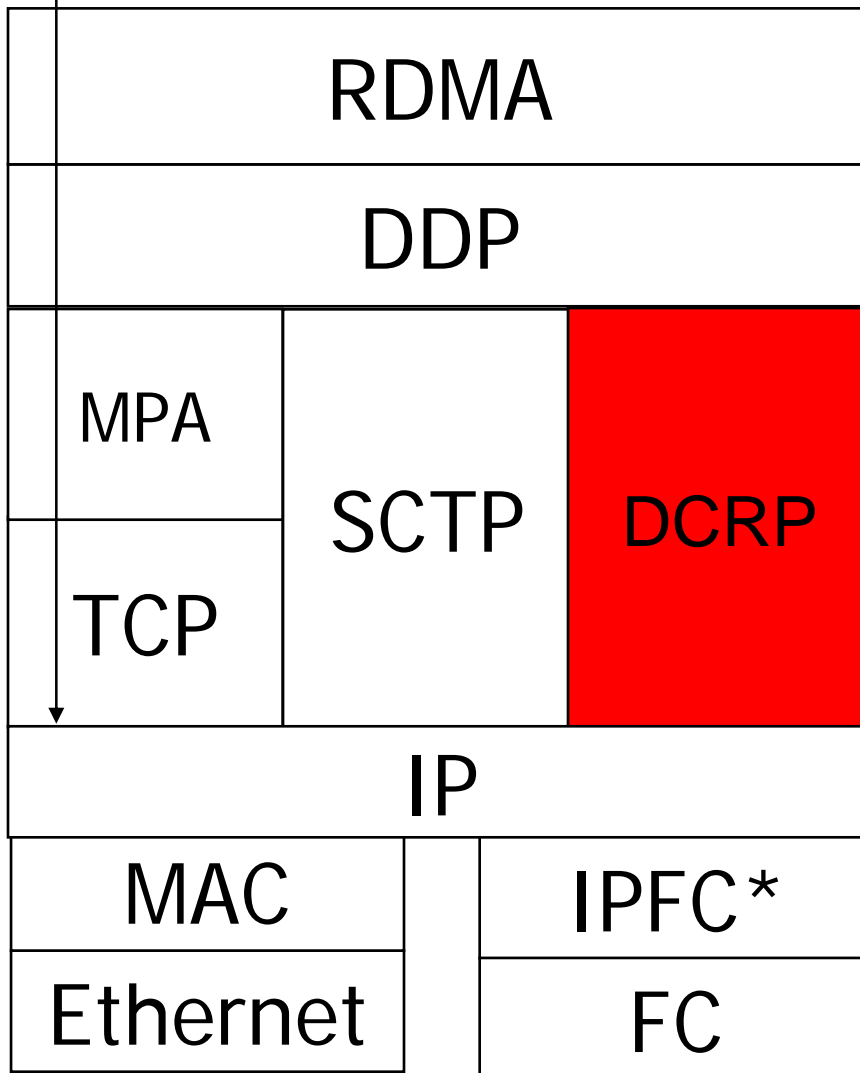
# Why TCP doesn't scale well

- TCP is an byte stream protocol while RDMA is a message protocol
  - Marker PDU Aligned Frame (MPA – RFC 5044) adds markers to enable the RDMA over TCP – See next slide
- Complex protocol handling
  - Difficult to separate data placement and protocol handling
  - Results in need for TCP Offload Engine (TOE)
- Large number of connections is burdensome for TCP
- Multi-homing limits optimal path



# DCRP Approach

iWARP

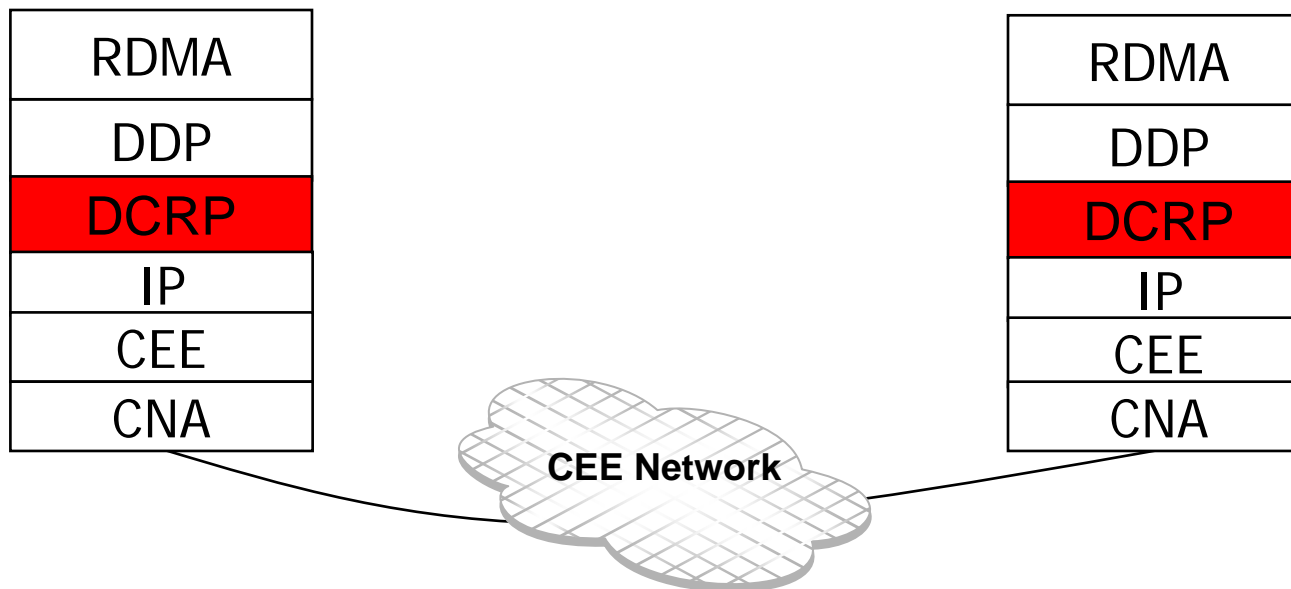


- DCRP (Data Center RDMA Protocol) is a simplified transport layer intended to operate on low loss networks
  - To be used instead of MPA/TCP or SCTP
- DCRP uses an IP header
  - Permits standard Discovery
    - Via DNS, ARP, ARP Cache, etc.
  - For compatibility with IPsec and ESP (Encapsulating Security Protocol)
  - Can be routed to other CEE subnets (special CEE handling for the ESP/IP protocol)
- Internet Protocol over Fibre Channel\* (IPFC\*) will be a streamlined version of IPFC so that no state information is required



# DCRP over Converged Enhanced Ethernet (CEE)

- DCRP would enable CNA to CNA communications over the CEE Network



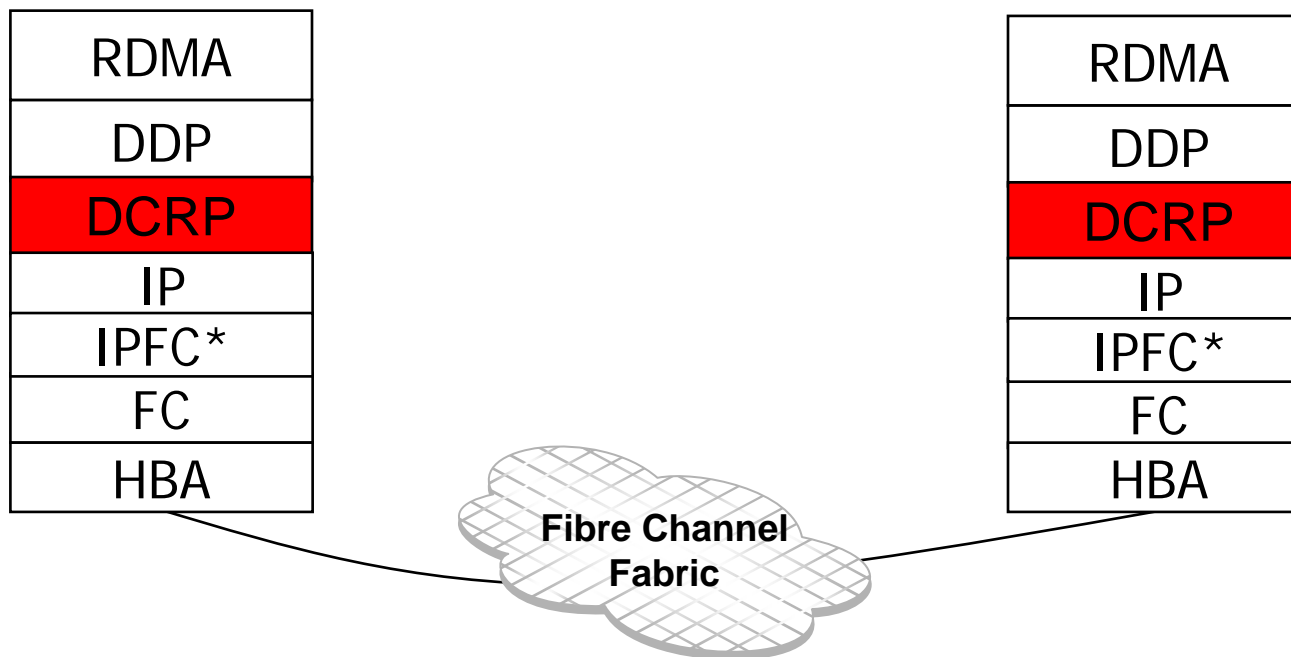
RDMA = Remote Direct Memory Access

DDP = Direct Data Placement

CNA = Converged Network Adapter

# DCRP over FC

- DCRP could be mapped to FC with IFCP



RDMA = Remote Direct Memory Access

DDP = Direct Data Placement

CNA = Converged Network Adapter

IPFC\* = Streamlined version of IPFC

# Where to standardize?

- IETF used to have the IP Storage Working Group that might have handled topics like DCRP. The WG was established to not work on:
  - Modifications to internet transport protocols or approaches requiring transport protocol options that are not widely supported, although the WG may recommend use of such options for block storage traffic.
- IETF mission statement as defined in RFC 3935:
  - The goal of the IETF is to make the Internet work better.
- DCRP is a new transport protocol that is only intended to be used in high performance computing clusters so the IETF is not appropriate



# Standardize in T11

- T11 is currently known as technical committee to define Fibre Channel Interfaces
- T11 has defined many more storage protocols in the past including HIPPI, IPI, SNPING, and various IETF MIBs
- Brocade would like to open up the breadth of the T11 committee to cover topics like DCRP
- Brocade would like to form a study group to pursue solutions to provide RDMA transport over CEE and FC that uses IP as a common point

