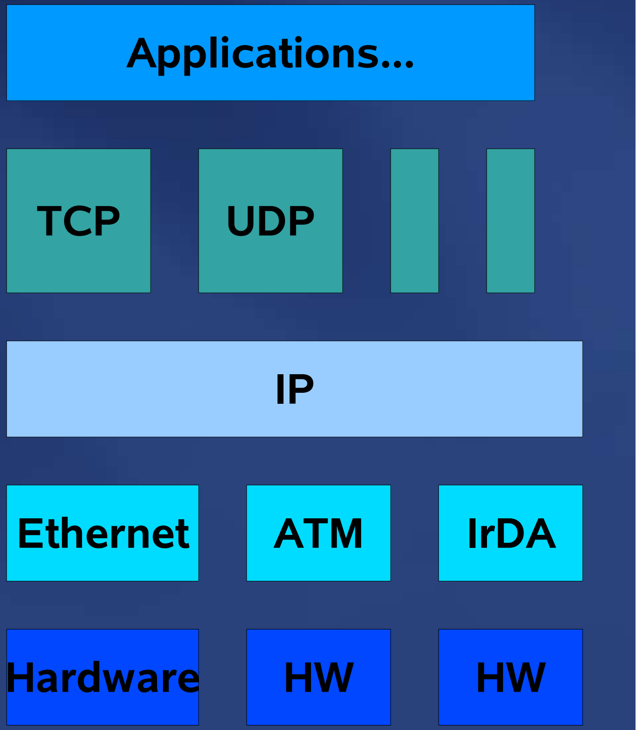
Infiniband intro

- The TCP/IP world: quick overview
- The Infiniband world
- IB today: market and political situation

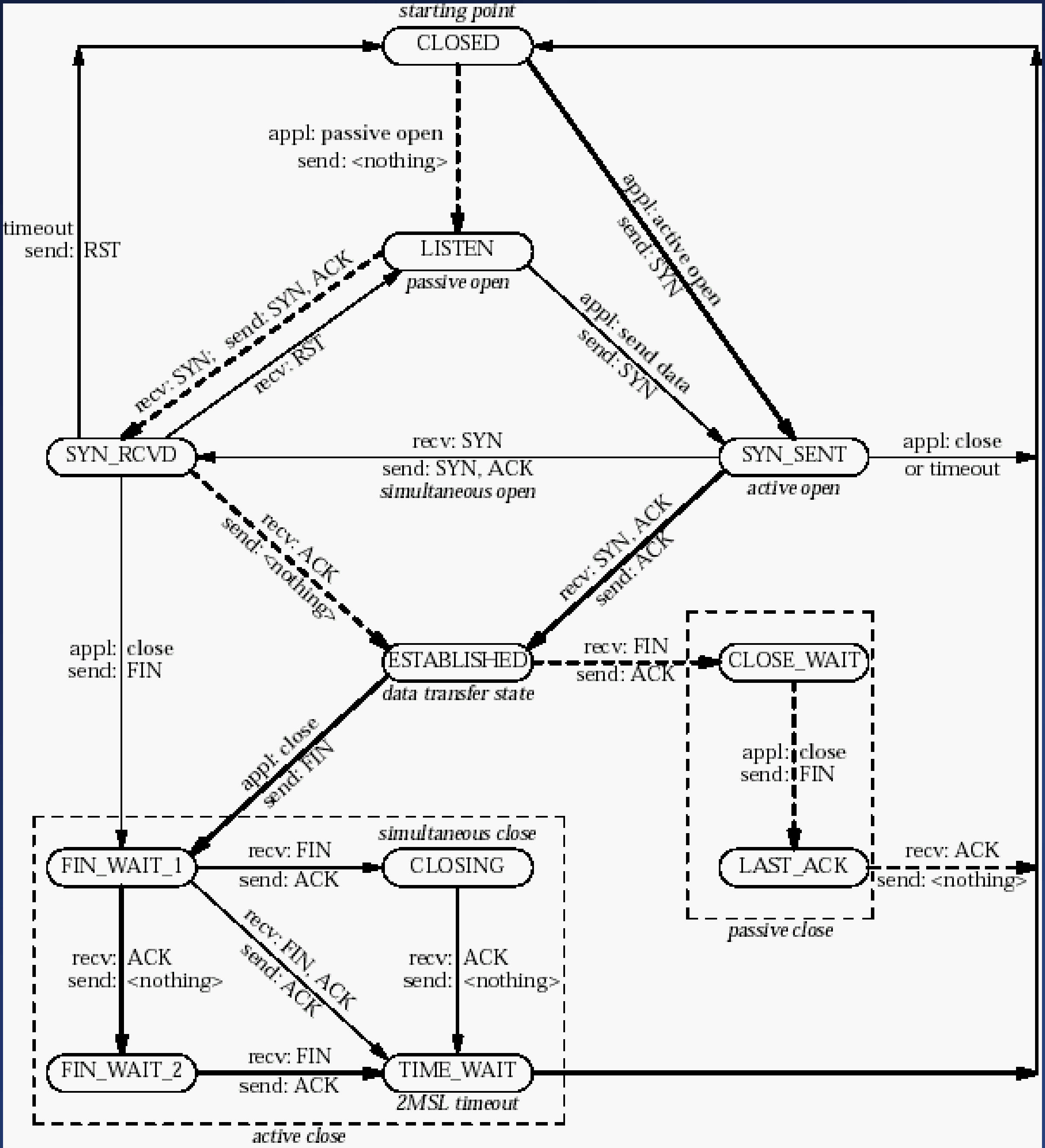
Internet protocol

- is an abstraction over different *data link* protocols
- *IP addresses* for *host identification*
- packet switching, unreliable
- several services on top:
  - UDP: connectionless lightweight transport
    - fast, no error correction
  - TCP: reliable transport AD ~1980
    - designed for low-bandwidth unreliable links
    - reliable, fault tolerant, socket semantics
  - many others:
    - ICMP (network control)
    - IPSec
    - SCTP...



TCP/IP(/Ethernet) problems

- performance scaling (in Ethernet and/or TCP)
  - memory copies, checksumming, ramp-up, packet size
  - error correction: throw away bad backets
  - more bandwidth -> more significant problems
  - hardware implementation very expensive (but: RDMA/Ethernet?)
- services (not) provided
  - QoS – very basic, no guarantees
  - load balancing, security (IPSec): complicated afterthoughts
  - address space too small (Ipv6? has its own problems)
  - fabric resilience: complicated (routing protocols on IP level or STP/link aggregation on Ethernet level)

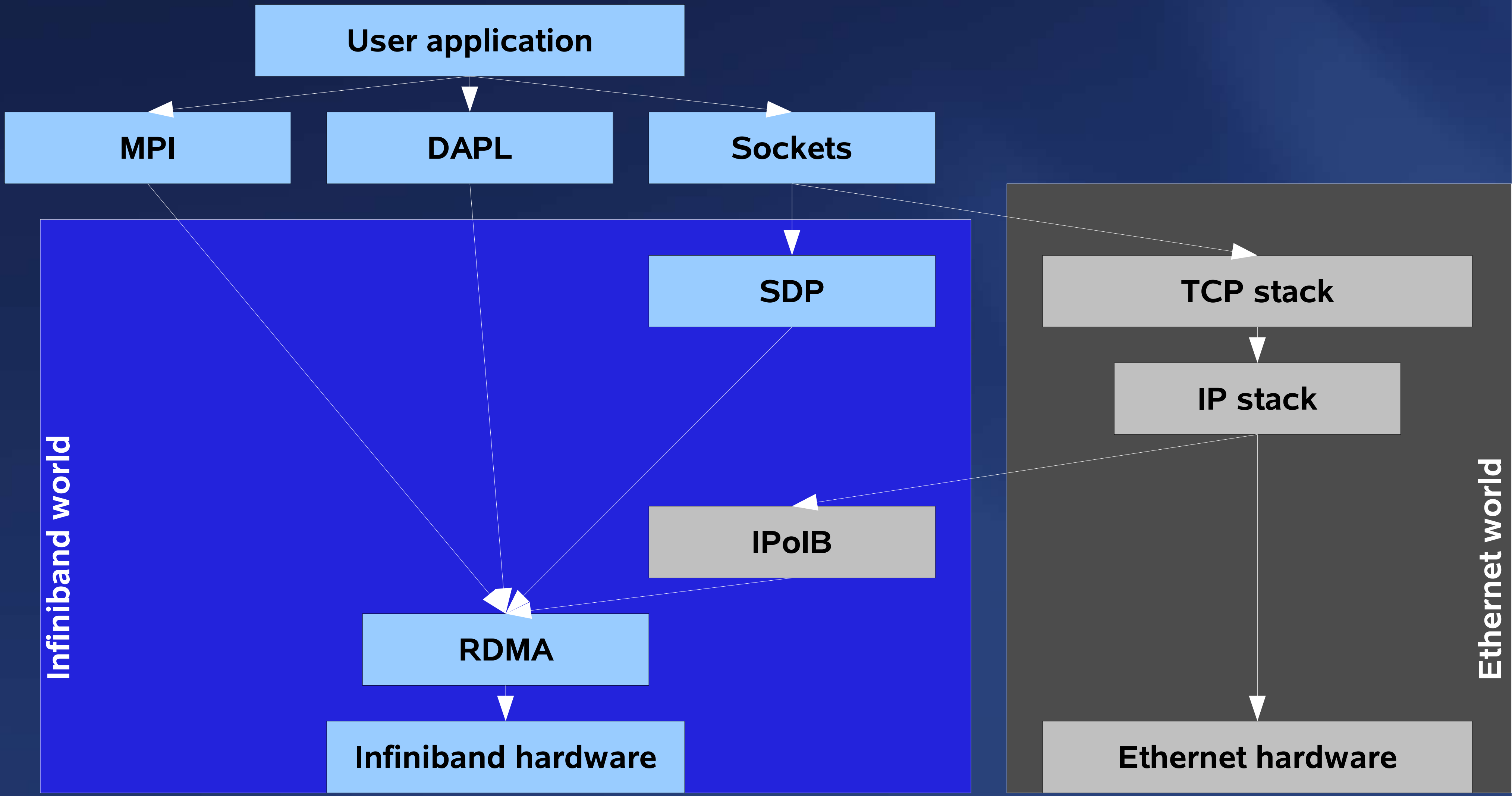


━━▶	*normal transitions for client*
- - ▶	*normal transitions for server*
appl:	*state transitions taken when application issues operation*
recv:	*state transitions taken when segment received*
send:	*what is sent for this transition*

Infiniband basics

- low-latency, high-bandwidth new interconnect
- a whole *new protocol stack* from hardware up
- IP addresses for host identification – but no IP protocol
- therefore, protocol conversion necessary to connect to other networks (Ethernet, Myrinet, Fibre Channel)
- *designed to be* implemented in / assisted by HW
- open standard backed by many companies
- opensource software: under development, also in 2.6.10-mm1+
- built-in QoS, link failover, fabric monitoring, load balancing...
- API: support for sockets, MPI, DAPL, SRP (iSER) ...
- Bandwidth/price is very good
- at CERN: native port of RFIO (CASTOR), basic tests

Network layers concept



Infiniband problems and politics

- One Chipmaker to Bind Them (Mellanox)
- Market slow, mainly MPI only (but: some supercomputers, VirginiaTech etc)
- Lack of expertise and experience
- Disruptive cabling, problems with long distance, external connectivity
- Drivers are complicated (memory management issues)
- No native storage products available
- API is only functionally defined in the standard –> started with vendor-specific implementations in closed source!
- OpenIB: standard API, opensource, but: is it too late?

Literature

TCP/IP

- RFC791 (IP), RFC793, STD7 (TCP) and others
- TCP/IP Illustrated (G.Wright, R. Stevens)
- google://Sally Floyd
- WAN data transfer papers by A. Hirstius

Infiniband

- http://www.buyya.com/superstorage/chap42.pdf
- http://www.infinibandta.org/
- http://www.openib.org/
- http://cern.ch/ahorvath/ib