

Networking & Systems Programming for HFT

Complete Guide: Protocols, Low-Latency Techniques, and Production Systems

Table of Contents

Part I: Network Fundamentals

1. OSI Model and TCP/IP Stack
2. Socket Programming Basics
3. TCP vs UDP Trade-offs
4. Network Performance Metrics

Part II: High-Performance Networking

5. Non-Blocking I/O and select/poll/epoll
6. Zero-Copy Techniques
7. Kernel Bypass (DPDK, RDMA)
8. Multicast and Market Data Distribution

Part III: HFT-Specific Optimizations

9. Ultra-Low Latency Network Design
10. FIX Protocol Optimization
11. Hardware Timestamping
12. Network Topology and Colocation

Chapter 1: Socket Programming Fundamentals

1.1 Foundational Overview

Layman's Intuition

Think of sockets as telephone connections. A server socket is like a business phone number—clients "call" (connect to) it. Once connected, both sides can talk (send/receive data) bi-directionally. Just as phone systems have protocols (how to dial, who speaks first), network sockets follow protocols like TCP and UDP.

Technical Foundation

Socket: Endpoint for network communication

Socket Address Structure:

- **IP Address:** Identifies machine (e.g., 192.168.1.100)
- **Port Number:** Identifies application (e.g., 8080)
- **Combined:** 192.168.1.100:8080 (socket address)

Socket Types:

1. SOCK_STREAM (TCP):

- Connection-oriented
- Reliable, ordered delivery
- Byte-stream (no message boundaries)
- Use: HTTP, FIX protocol, database connections

2. SOCK_DGRAM (UDP):

- Connectionless
- Unreliable, unordered
- Message-oriented (datagram boundaries preserved)
- Use: Market data multicast, DNS, streaming

Industry Relevance

HFT Systems:

- **Market Data:** UDP multicast for low-latency tick distribution
- **Order Entry:** TCP for reliable order submission
- **Colocation:** Direct exchange connections (microsecond latency)

Backend Systems:

- **HTTP/REST APIs:** TCP for web services
- **Microservices:** gRPC, Thrift over TCP
- **Databases:** TCP connections for queries

1.2 TCP Socket Programming

Server-Client Model

TCP Server (Basic):

```
#include <sys/socket.h>;
#include <netinet/in.h>;
#include <arpa/inet.h>;
#include <unistd.h>;
#include <cstring>;
#include <iostream>;

class TCPServer {
private:
    int server_fd;
    int port;

public:
    TCPServer(int port) : port(port), server_fd(-1) {}

    bool start() {
        // Step 1: Create socket
        server_fd = socket(AF_INET, SOCK_STREAM, 0);
        if (server_fd < 0) {
            perror("socket creation failed");
            return false;
        }

        // Step 2: Set socket options
        int opt = 1;
        if (setsockopt(server_fd, SOL_SOCKET, SO_REUSEADDR,
            &opt, sizeof(opt)) < 0) {
            perror("setsockopt failed");
            return false;
        }

        // Step 3: Bind to address
        sockaddr_in address;
        memset(&address, 0, sizeof(address));
        address.sin_family = AF_INET;
        address.sin_addr.s_addr = INADDR_ANY; // Accept all interfaces
        address.sin_port = htons(port);

        if (bind(server_fd, (sockaddr*)&address, sizeof(address)) < 0) {
            perror("bind failed");
            return false;
        }

        // Step 4: Listen for connections
        if (listen(server_fd, 10) < 0) { // Backlog = 10
            perror("listen failed");
            return false;
        }

        std::cout <<< "Server listening on port " <<< port <<< "\n";
        return true;
    }
}
```

```

void run() {
    while (true) {
        // Step 5: Accept client connection
        sockaddr_in client_addr;
        socklen_t client_len = sizeof(client_addr);

        int client_fd = accept(server_fd,
                               (sockaddr*)&client_addr,
                               &client_len);

        if (client_fd < 0) {
            perror("accept failed");
            continue;
        }

        // Get client IP
        char client_ip[INET_ADDRSTRLEN];
        inet_ntop(AF_INET, &client_addr.sin_addr,
                  client_ip, INET_ADDRSTRLEN);

        std::cout <<< "Client connected from " <<< client_ip
                    <<< ":" <<< ntohs(client_addr.sin_port) <<< "\n";

        // Step 6: Handle client (could spawn thread here)
        handle_client(client_fd);

        close(client_fd);
    }
}

void handle_client(int client_fd) {
    char buffer[1024];

    while (true) {
        // Receive data
        ssize_t bytes_read = recv(client_fd, buffer,
                                   sizeof(buffer) - 1, 0);

        if (bytes_read <= 0) {
            if (bytes_read == 0) {
                std::cout <<< "Client disconnected\n";
            } else {
                perror("recv failed");
            }
            break;
        }

        buffer[bytes_read] = '\0';
        std::cout <<< "Received: " <<< buffer <<< "\n";

        // Echo back to client
        send(client_fd, buffer, bytes_read, 0);
    }
}

```

```

~TCPServer() {
    if (server_fd >= 0) {
        close(server_fd);
    }
}

};

int main() {
    TCPServer server(8080);

    if (server.start()) {
        server.run();
    }

    return 0;
}

```

TCP Client:

```

class TCPClient {
private:
    int sock_fd;
    std::string server_ip;
    int server_port;

public:
    TCPClient(const std::string& ip, int port)
        : server_ip(ip), server_port(port), sock_fd(-1) {}

    bool connect() {
        // Create socket
        sock_fd = socket(AF_INET, SOCK_STREAM, 0);
        if (sock_fd < 0) {
            perror("socket creation failed");
            return false;
        }

        // Server address
        sockaddr_in server_addr;
        memset(&server_addr, 0, sizeof(server_addr));
        server_addr.sin_family = AF_INET;
        server_addr.sin_port = htons(server_port);

        // Convert IP address
        if (inet_pton(AF_INET, server_ip.c_str(),
                     &server_addr.sin_addr) <= 0) {
            perror("invalid address");
            return false;
        }

        // Connect to server
        if (::connect(sock_fd, (sockaddr*)&server_addr,
                     sizeof(server_addr)) < 0) {
            perror("connection failed");
            return false;
        }
    }
}

```

```

    }

    std::cout <<< "Connected to " <<< server_ip
                <<< ":" <<< server_port <<< "\n";
    return true;
}

bool send_message(const std::string& message) {
    ssize_t bytes_sent = send(sock_fd, message.c_str(),
                              message.length(), 0);

    if (bytes_sent < 0) {
        perror("send failed");
        return false;
    }

    return true;
}

std::string receive_message() {
    char buffer[1024];
    ssize_t bytes_read = recv(sock_fd, buffer,
                              sizeof(buffer) - 1, 0);

    if (bytes_read <= 0) {
        return "";
    }

    buffer[bytes_read] = '\0';
    return std::string(buffer);
}

~TCPClient() {
    if (sock_fd >= 0) {
        close(sock_fd);
    }
}

};

// Usage
int main() {
    TCPClient client("127.0.0.1", 8080);

    if (!client.connect()) {
        return 1;
    }

    // Send message
    client.send_message("Hello, server!");

    // Receive echo
    std::string response = client.receive_message();
    std::cout <<< "Received: " <<< response <<< "\n";

    return 0;
}

```

Multithreaded TCP Server

Thread-per-Connection Model:

```
#include <thread>
#include <vector>

class MultithreadedTCPServer {
private:
    int server_fd;
    int port;
    std::vector<std::thread> client_threads;
    std::atomic<bool> running{true};

public:
    MultithreadedTCPServer(int port) : port(port), server_fd(-1) {}

    bool start() {
        server_fd = socket(AF_INET, SOCK_STREAM, 0);
        if (server_fd < 0) return false;

        int opt = 1;
        setsockopt(server_fd, SOL_SOCKET, SO_REUSEADDR, &opt, sizeof(opt));

        sockaddr_in address;
        memset(&address, 0, sizeof(address));
        address.sin_family = AF_INET;
        address.sin_addr.s_addr = INADDR_ANY;
        address.sin_port = htons(port);

        if (bind(server_fd, (sockaddr*)&address, sizeof(address)) < 0) {
            return false;
        }

        if (listen(server_fd, 100) < 0) { // Higher backlog for concurrent
            return false;
        }

        return true;
    }

    void run() {
        while (running) {
            sockaddr_in client_addr;
            socklen_t client_len = sizeof(client_addr);

            int client_fd = accept(server_fd, (sockaddr*)&client_addr,
                                   &client_len);

            if (client_fd < 0) {
                continue;
            }

            // Spawn thread for this client
            client_threads.emplace_back([this, client_fd]() {
                handle_client(client_fd);
            });
        }
    }
};
```

```

        });
    }
}

void handle_client(int client_fd) {
    // Per-thread buffer (thread-safe)
    char buffer[4096];

    while (running) {
        ssize_t bytes_read = recv(client_fd, buffer,
                                   sizeof(buffer), 0);

        if (bytes_read <= 0) break;

        // Process request
        process_request(buffer, bytes_read);

        // Send response
        send(client_fd, buffer, bytes_read, 0);
    }

    close(client_fd);
}

void process_request(char* data, size_t len) {
    // Application logic here
}

void shutdown() {
    running = false;

    for (auto& t : client_threads) {
        if (t.joinable()) {
            t.join();
        }
    }

    close(server_fd);
}
};

```

1.3 UDP Socket Programming

UDP Server

```

class UDPServer {
private:
    int sock_fd;
    int port;

public:
    UDPServer(int port) : port(port), sock_fd(-1) {}

```



```

bool start() {
    // Create UDP socket
    sock_fd = socket(AF_INET, SOCK_DGRAM, 0);
    if (sock_fd < 0) {
        perror("socket creation failed");
        return false;
    }

    // Bind to address
    sockaddr_in server_addr;
    memset(&server_addr, 0, sizeof(server_addr));
    server_addr.sin_family = AF_INET;
    server_addr.sin_addr.s_addr = INADDR_ANY;
    server_addr.sin_port = htons(port);

    if (bind(sock_fd, (sockaddr*)&server_addr,
            sizeof(server_addr)) < 0) {
        perror("bind failed");
        return false;
    }

    std::cout <<< "UDP server listening on port " <<< port <<< "\n";
    return true;
}

void run() {
    char buffer[65536]; // Max UDP datagram size

    while (true) {
        sockaddr_in client_addr;
        socklen_t client_len = sizeof(client_addr);

        // Receive datagram
        ssize_t bytes_read = recvfrom(sock_fd, buffer,
                                       sizeof(buffer), 0,
                                       (sockaddr*)&client_addr,
                                       &client_len);

        if (bytes_read < 0) {
            perror("recvfrom failed");
            continue;
        }

        buffer[bytes_read] = '\0';

        char client_ip[INET_ADDRSTRLEN];
        inet_ntop(AF_INET, &client_addr.sin_addr,
                  client_ip, INET_ADDRSTRLEN);

        std::cout <<< "Received " <<< bytes_read <<< " bytes from "
                  <<< client_ip <<< ":" <<< ntohs(client_addr.sin_port)
                  <<< "\n";

        // Echo back
        sendto(sock_fd, buffer, bytes_read, 0,
               (sockaddr*)&client_addr, client_len);
    }
}

```

```

    }
}

~UDPServer() {
    if (sock_fd >= 0) {
        close(sock_fd);
    }
}

};

```

UDP Client

```

class UDPClient {
private:
    int sock_fd;
    sockaddr_in server_addr;

public:
    UDPClient(const std::string& server_ip, int server_port) : sock_fd(-1) {
        sock_fd = socket(AF_INET, SOCK_DGRAM, 0);

        memset(&server_addr, 0, sizeof(server_addr));
        server_addr.sin_family = AF_INET;
        server_addr.sin_port = htons(server_port);
        inet_pton(AF_INET, server_ip.c_str(), &server_addr.sin_addr);
    }

    void send_message(const std::string& message) {
        sendto(sock_fd, message.c_str(), message.length(), 0,
            (sockaddr*)&server_addr, sizeof(server_addr));
    }

    std::string receive_message() {
        char buffer[65536];
        socklen_t server_len = sizeof(server_addr);

        ssize_t bytes_read = recvfrom(sock_fd, buffer, sizeof(buffer), 0,
            (sockaddr*)&server_addr, &server_len);

        if (bytes_read < 0) {
            return "";
        }

        buffer[bytes_read] = '\0';
        return std::string(buffer);
    }

    ~UDPClient() {
        if (sock_fd >= 0) {
            close(sock_fd);
        }
    }
};

```

1.4 UDP Multicast for Market Data

Layman's Intuition

Multicast is like a radio broadcast. The sender transmits once, and multiple receivers tune in to the same "channel" (multicast group). Efficient for distributing identical data (market ticks) to many clients simultaneously.

Multicast Sender

```
class MulticastSender {
private:
    int sock_fd;
    sockaddr_in multicast_addr;

public:
    MulticastSender(const std::string& multicast_ip, int port) : sock_fd(-1) {
        // Create UDP socket
        sock_fd = socket(AF_INET, SOCK_DGRAM, 0);
        if (sock_fd <= 0) {
            throw std::runtime_error("socket creation failed");
        }

        // Set multicast address
        memset(&multicast_addr, 0, sizeof(multicast_addr));
        multicast_addr.sin_family = AF_INET;
        multicast_addr.sin_port = htons(port);
        inet_pton(AF_INET, multicast_ip.c_str(), &multicast_addr.sin_addr);

        // Set multicast TTL (Time To Live)
        unsigned char ttl = 64;
        setsockopt(sock_fd, IPPROTO_IP, IP_MULTICAST_TTL, &ttl, sizeof(ttl));

        // Disable loopback (don't receive own messages)
        unsigned char loop = 0;
        setsockopt(sock_fd, IPPROTO_IP, IP_MULTICAST_LOOP, &loop, sizeof(loop));
    }

    void send(const void* data, size_t len) {
        sendto(sock_fd, data, len, 0,
            (sockaddr*)&multicast_addr, sizeof(multicast_addr));
    }

    ~MulticastSender() {
        if (sock_fd >= 0) {
            close(sock_fd);
        }
    }
};
```

Multicast Receiver

```
class MulticastReceiver {
private:
    int sock_fd;

public:
    MulticastReceiver(const std::string& multicast_ip, int port,
                      const std::string& interface_ip = "0.0.0.0") {
        // Create UDP socket
        sock_fd = socket(AF_INET, SOCK_DGRAM, 0);
        if (sock_fd < 0) {
            throw std::runtime_error("socket creation failed");
        }

        // Allow multiple sockets to bind to same port
        int reuse = 1;
        setsockopt(sock_fd, SOL_SOCKET, SO_REUSEADDR, &reuse, sizeof(reuse));

        // Bind to multicast port
        sockaddr_in local_addr;
        memset(&local_addr, 0, sizeof(local_addr));
        local_addr.sin_family = AF_INET;
        local_addr.sin_addr.s_addr = INADDR_ANY;
        local_addr.sin_port = htons(port);

        if (bind(sock_fd, (sockaddr*)&local_addr, sizeof(local_addr)) < 0) {
            throw std::runtime_error("bind failed");
        }

        // Join multicast group
        ip_mreq mreq;
        inet_pton(AF_INET, multicast_ip.c_str(), &mreq.imr_multicast);
        inet_pton(AF_INET, interface_ip.c_str(), &mreq.imr_interface);

        if (setsockopt(sock_fd, IPPROTO_IP, IP_ADD_MEMBERSHIP,
                       &mreq, sizeof(mreq)) < 0) {
            throw std::runtime_error("join multicast group failed");
        }
    }

    ssize_t receive(void* buffer, size_t len) {
        return recvfrom(sock_fd, buffer, len, 0, nullptr, nullptr);
    }

    ~MulticastReceiver() {
        if (sock_fd >= 0) {
            close(sock_fd);
        }
    }
};
```

HFT Market Data Example

```
struct MarketTick {
    uint64_t timestamp;
    uint32_t symbol_id;
    double price;
    uint32_t volume;
} __attribute__((packed)); // No padding for network transmission

// Exchange multicast sender
void exchange_feed_publisher() {
    MulticastSender sender("239.255.0.1", 9000);

    while (true) {
        MarketTick tick;
        tick.timestamp = get_timestamp_ns();
        tick.symbol_id = 1;
        tick.price = 100.50;
        tick.volume = 1000;

        sender.send(&tick, sizeof(tick));

        usleep(1000); // 1ms between ticks
    }
}

// HFT client receiver
void market_data_receiver() {
    MulticastReceiver receiver("239.255.0.1", 9000);

    MarketTick tick;

    while (true) {
        ssize_t bytes = receiver.receive(&tick, sizeof(tick));

        if (bytes == sizeof(tick)) {
            uint64_t recv_time = get_timestamp_ns();
            uint64_t latency = recv_time - tick.timestamp;

            std::cout <<< "Latency: " <<< latency <<< " ns\n";

            // Process tick
            on_market_tick(tick);
        }
    }
}
```

1.5 Non-Blocking I/O with epoll

Layman's Intuition

Blocking I/O is like waiting at a checkout line—you can't do anything else. Non-blocking I/O with epoll is like having a pager system: you give the cashier your pager, shop for more items, and get notified when it's your turn. One thread can monitor thousands of connections efficiently.

epoll Basics

epoll vs select/poll:

Feature	select	poll	epoll
Max FDs	1024	Unlimited	Unlimited
Performance	O(n)	O(n)	O(1)
Modification	Copy all	Copy all	Incremental
Use Case	Legacy	Moderate	High-performance

epoll Server Implementation

```
#include <sys/epoll.h>
#include <fcntl.h>

class EpollServer {
private:
    int server_fd;
    int epoll_fd;
    static constexpr int MAX_EVENTS = 1024;

public:
    EpollServer(int port) {
        // Create server socket
        server_fd = socket(AF_INET, SOCK_STREAM, 0);

        // Set non-blocking
        set_non_blocking(server_fd);

        // Bind and listen
        sockaddr_in addr;
        memset(&addr, 0, sizeof(addr));
        addr.sin_family = AF_INET;
        addr.sin_addr.s_addr = INADDR_ANY;
        addr.sin_port = htons(port);

        int opt = 1;
        setsockopt(server_fd, SOL_SOCKET, SO_REUSEADDR, &opt, sizeof(opt));

        bind(server_fd, (sockaddr*)&addr, sizeof(addr));
        listen(server_fd, 128);
    }
};
```

```

    // Create epoll instance
    epoll_fd = epoll_create1(0);
    if (epoll_fd < 0) {
        throw std::runtime_error("epoll_create1 failed");
    }

    // Add server socket to epoll
    epoll_event ev;
    ev.events = EPOLLIN; // Interested in read events
    ev.data.fd = server_fd;

    epoll_ctl(epoll_fd, EPOLL_CTL_ADD, server_fd, &ev);
}

void set_non_blocking(int fd) {
    int flags = fcntl(fd, F_GETFL, 0);
    fcntl(fd, F_SETFL, flags | O_NONBLOCK);
}

void run() {
    epoll_event events[MAX_EVENTS];

    while (true) {
        // Wait for events (timeout = -1 means infinite)
        int num_events = epoll_wait(epoll_fd, events, MAX_EVENTS, -1);

        for (int i = 0; i < num_events; i++) {
            int fd = events[i].data.fd;

            if (fd == server_fd) {
                // New connection
                handle_new_connection();
            } else {
                // Data from existing connection
                handle_client_data(fd, events[i].events);
            }
        }
    }
}

void handle_new_connection() {
    while (true) {
        sockaddr_in client_addr;
        socklen_t client_len = sizeof(client_addr);

        int client_fd = accept(server_fd, (sockaddr*)&client_addr,
                               &client_len);

        if (client_fd < 0) {
            if (errno == EAGAIN || errno == EWOULDBLOCK) {
                // No more connections to accept
                break;
            }
            perror("accept failed");
            break;
        }
    }
}

```

```

        // Set client socket non-blocking
        set_non_blocking(client_fd);

        // Add to epoll
        epoll_event ev;
        ev.events = EPOLLIN | EPOLLET; // Edge-triggered mode
        ev.data.fd = client_fd;

        epoll_ctl(epoll_fd, EPOLL_CTL_ADD, client_fd, &ev);

        std::cout <<< "New client connected: fd=" <<< client_fd <<< "\n";
    }
}

void handle_client_data(int client_fd, uint32_t events) {
    if (events & EPOLLIN) {
        // Data available to read
        char buffer[4096];

        while (true) {
            ssize_t bytes_read = recv(client_fd, buffer,
                                      sizeof(buffer), 0);

            if (bytes_read < 0) {
                if (errno == EAGAIN || errno == EWOULDBLOCK) {
                    // No more data to read
                    break;
                }
                // Error
                close_connection(client_fd);
                break;
            }

            if (bytes_read == 0) {
                // Client disconnected
                close_connection(client_fd);
                break;
            }

            // Process data
            process_data(client_fd, buffer, bytes_read);
        }
    }

    if (events & (EPOLLHUP | EPOLLERR)) {
        // Connection error or hangup
        close_connection(client_fd);
    }
}

void process_data(int client_fd, const char* data, size_t len) {
    // Application logic
    // Echo back for demo
    send(client_fd, data, len, MSG_NOSIGNAL);
}

```



```

void close_connection(int client_fd) {
    epoll_ctl(epoll_fd, EPOLL_CTL_DEL, client_fd, nullptr);
    close(client_fd);
    std::cout <<< "Client disconnected: fd=" <<< client_fd <<< "\n";
}

~EpollServer() {
    close(epoll_fd);
    close(server_fd);
}

};

// Usage
int main() {
    EpollServer server(8080);
    server.run();
    return 0;
}

```

Key Epoll Concepts:

Edge-Triggered (ET) vs Level-Triggered (LT):

- **Level-Triggered (default):** `epoll_wait` returns whenever data is available
 - Easier to use
 - May return multiple times for same event
- **Edge-Triggered (EPOLLET):** `epoll_wait` returns only when state changes
 - More efficient
 - Must read all data (loop until `EAGAIN`)
 - HFT: Prefer ET for lower overhead

1.6 Zero-Copy Techniques

`sendfile()` System Call

Traditional Copy:

```

Disk → Kernel Buffer → User Buffer → Socket Buffer → NIC
      (copy 1)         (copy 2)         (copy 3)

```

Zero-Copy with `sendfile()`:

```

Disk → Kernel Buffer → NIC
      (DMA transfer, no CPU copies)

```

Implementation:

```

#include <sys/sendfile.h>;
#include <sys/stat.h>;

void send_file_traditional(int client_fd, const char* filename) {
    int file_fd = open(filename, O_RDONLY);
    if (file_fd < 0) return;

    char buffer[8192];
    ssize_t bytes_read;

    while ((bytes_read = read(file_fd, buffer, sizeof(buffer))) > 0) {
        send(client_fd, buffer, bytes_read, 0);
    }

    close(file_fd);
}

void send_file_zero_copy(int client_fd, const char* filename) {
    int file_fd = open(filename, O_RDONLY);
    if (file_fd < 0) return;

    struct stat stat_buf;
    fstat(file_fd, &stat_buf);

    // Zero-copy transfer: file → socket
    sendfile(client_fd, file_fd, nullptr, stat_buf.st_size);

    close(file_fd);
}

```

Performance Comparison:

```

void benchmark_file_transfer() {
    // Create 100MB test file
    create_test_file("test.dat", 100 * 1024 * 1024);

    int client = create_client_socket();

    // Traditional
    auto start = std::chrono::high_resolution_clock::now();
    send_file_traditional(client, "test.dat");
    auto end = std::chrono::high_resolution_clock::now();

    auto traditional_time = std::chrono::duration_cast<std::chrono::milliseconds>(
        (end - start).count());

    // Zero-copy
    start = std::chrono::high_resolution_clock::now();
    send_file_zero_copy(client, "test.dat");
    end = std::chrono::high_resolution_clock::now();

    auto zerocopy_time = std::chrono::duration_cast<std::chrono::milliseconds>(
        (end - start).count());
}

```

```
std::cout &&& "Traditional: " &&& traditional_time &&& " ms\n";  
std::cout &&& "Zero-copy: " &&& zerocopy_time &&& " ms\n";  
std::cout &&& "Speedup: " &&& traditional_time / (double)zerocopy_time &&&  
}
```

This comprehensive textbook continues with kernel bypass techniques (DPDK, RDMA), FIX protocol optimization, hardware timestamping, and complete HFT network architecture examples. Each section provides production-ready code, performance benchmarks, and real-world deployment strategies.