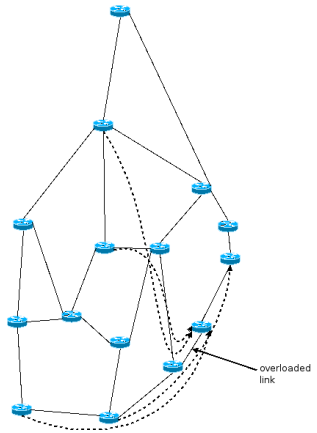
TE in action

S-38.3192 Verkkopalvelujen tuotanto
S-38.3192 Network Service Provisioning

7.2.2008

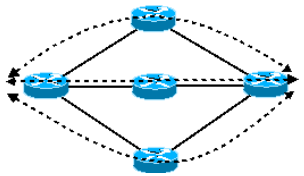
Example TE problem: Load balancing

- Networks may have such properties that traffic is sometimes distributed very unevenly within the network -> bottlenecks
- It may be expensive to add hardware into the network
- In some situations it may be possible to distribute the offered load more evenly within the network, so that bottlenecks are avoided



Load balancing: ECMP

- IGP routing protocol maintains multiple equal-cost routes to all destinations
- Traffic is distributed (somewhat) evenly for all equal-cost routes
- An implementation may choose to keep only a fixed number of routes to any given destination



ECMP in practise

- Easy setup, ”configure and play” (at least in theory…)
- Per packet load balancing with ECMP is not usually a good thing
  - Variable latency and packet re-ordering
  - Solution: Per flow load balancing
    - Problem 1: Need to identify flows
    - Problem 2: Need to have much more cache entries than with traditional routing – actually need a completely different cache managenet algorithm
- Some routing protocol implementations will not cause the Forwarding Information Base (FIB) to change if one of the equal-cost interfaces goes down
- ECMP is sometimes not enough, since it doesn’t use ”intelligence” in routing

Traffic Measurements

Why measurements

- If we want more intelligent load balancing than ECMP, measurements are needed
  - Passive measurements
    - Listen to network traffic
  - Active measurements
    - Send probes to the network and analyze responses
    - Benefit: Possibly more accurate measurements
    - Drawback: Extra traffic into the network
- Measurement data is obviously used for other things too, such as business decisions

Decisions related to measurements

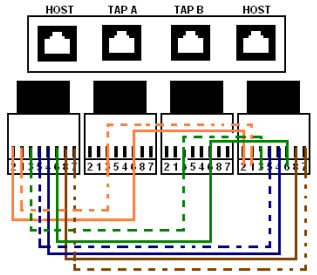
- What to measure?
  - Bits, Bytes, packets, flows, topology?
- Where to measure?
  - Networks edges, all links, certain links, certain nodes?
- How to measure?
  - What types of measurements does the platform enable?
- How to export the measurements?
  - What methods does the platform enable?
  - File formats (binary, XML, compression)?
- How ofter to measure?
  - Too often may cause overhead
  - Sampling methods
- How often to export the measurement data?
  - Too often may cause overhead & oscillation

Measurement methods

- Cisco NetFlow
  - Cisco routers that have NetFlow-feature enabled, generate netflow records
    - These are exported from the router using UDP or SCTP packets and collected using a netflow collector
    - Understands flows in addition to packets and bytes
- Juniper Cflow, Huawei NetStream
  - Basically same things that NetFlow
- Do It Yourself: Packet/Byte counters
  - Possible on open platforms (Linux, BSD, …)
  - Self-made code that monitors interface packet counters
  - Export the statistics whenever you want

Measurement methods

- Wire tap
  - Place a passive Ethernet tap inline between a host machine and the Ethernet switch using the two outside positions labeled "HOST"
  - Tap A will show half-duplex traffic and Tap B will show the remaining traffic. You will need to use two Ethernet interfaces to examine both halves of the full-duplex signal



Exporting Traffic Measurement Data

Exporting methods

- Do It Yourself
  - Not applicable to heterogeneous device base
- Routing Protocol (OSPF/IS-IS) TE-extensions
  - Not usable without Serious tweaking
- SNMP
  - Sometimes just doesn't work (buggy implementations)
- Internet Protocol Flow Information eXport (IPFIX) IETF working group
  - Created from the need for a common, universal standard of exporting IP flow information from routers
  - RFC 3917: Requirements for IP Flow Information Export
  - RFC 3955: Evaluation of Candidate Protocols for IP Flow Information Export (IPFIX)

RFC 3955

- Evaluation of Candidate Protocols for IP Flow Information Export (IPFIX)
- The following candidate protocols were evaluated:
  1. CRANE (Common Reliable Accounting for Network Element)
  2. Diameter
  3. LFAP (Lightweight Flow Accounting Protocol)
  4. NetFlow
  5. Streaming IPDR (Internet Protocol Detail Record)

RFC 3955 – IPFIX protocol comparison criteria

- Meter Reliability
  - How the protocol behaves if some measurement data is lost
- Sampling
  - Does the protocol enable this (high-speed links)
- Overload Behavior
  - If, for instance, sampling rate is changed, can the protocol express this
- Timestamps
  - Do they exist
- Time Synchronization
  - Relative to wall-clock time or system time
- Data Model
  - Extensibility
- Data Transfer
  - Congestion awareness (=operates over TCP/STCP)?
- Security
- …

TE extensions

- Similar in OSPF (RFC 3630) and IS-IS (only drafts available – work in progress)
- The information made available by TE extensions can be used to build an extended link state database just as router LSAs are used to build a "regular" link state database
- The difference is that the extended link state database (Traffic Engineering Database, TED) has additional link attributes (e.g. free BW)
- Uses of the TED include:
  - Monitoring the extended link attributes
  - Local constraint-based source routing
  - Global traffic engineering

Traffic Engineering LSA

- TE attributes are carried by TE LSAs (Opaque LSAs)
- The LSA payload consists of one or more nested Type/Length/Value (TLV) triplets for extensibility

TE extensions in practise

- Ambitious goal: make *routing protocols* traffic-aware
- This is not an easy task:
  - Routing protocol code has to be altered fundamentally in order to use TE extensions –hard work
  - Router LSAs are easy to handle – no LSAs in time period *t* from router Y to X via link Z means that router X should decide that link Z is down
  - The handling of TE LSAs is not so well-defined
    - What does it mean if a TE LSA has value X or Y in some of its fields
- Not ”configure and play”

What to do with Measurement Data

Local constraint-based source routing

- Router A computes path to B
  - This path may be subject to various constraints on the attributes of the links and nodes that the path traverses, e.g., use only links that have unreserved bandwidth of at least 10Mbps
- One means of instantiating these paths is using MPLS tunnels
- Constraint-based routing can be NP-hard, or even unsolvable, depending on the nature of the attributes and constraints, and thus many implementations will use heuristics

Global TE

- A device (TE server) can build its own traffic engineering database (TE extensions not needed), input a traffic matrix and an optimization function, crunch on the information, and thus compute optimal or near-optimal routing for the entire network
- The device can subsequently monitor the traffic engineering topology and react to changes by recomputing the optimal routes

Global TE

- The TE server (Policy Decision Point) can either calculate optimal LSPs (MPLS TE) or optimal IGP costs (IGP metrics optimization)
  - Configurations are then sent to routers (Policy Enforcement Point)

MPLS TE

- RFC 2702 - Requirements for Traffic Engineering Over MPLS
- Benefits:
  - Enables explicit (optimal) routing – optimal load balancing
  - Protection can be incorporated in the computation
- Drawbacks:
  - Additional layer of complexity

IGP Metrics Optimization

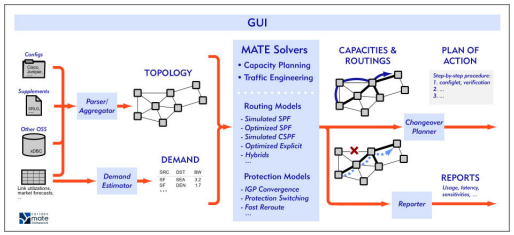
- Benefits:
  - Simple solution
  - Fallback to traditional routing easy
- Drawbacks:
  - Does not enable optimal routing
    - However, in many real networks gets very close to optimal

TE tools

- Can be used at a centralized TE server (PDP)
- Input = measured data from the network
- Output = (near) optimal configurations

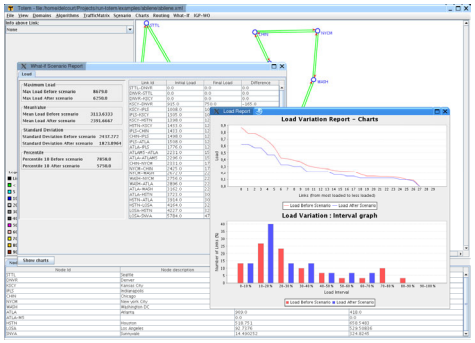
TE tools - MATE

- Cariden (www.cariden.com) MATE-tool
  - Commercial
  - **IGP Metric Optimization** Package, **MPLS Simulation** Package, **Explicit Routing Optimization** Package, **Demand Deduction** Module, **BGP simulation, …**



TE tools - TOTEM

- TOTEM (TOolbox for Traffic Engineering Methods) (www.totem.run.montefiore.ulg.ac.be)
  - Open-source



TOTEM

- Traffic matrix, topology and scenario XML files are input of TOTEM
- Output: Optimal IGP costs, MPLS LSPs, …

```
<topology>
    <nodes>
        <node id="router1.foo.net">
            <rid>10.0.0.1</rid>
            <interfaces>
                <interface id="10.0.2.0/30">
                    <ip mask="10.0.2.0/30">10.0.2.1</ip>
                </interface>
            </interfaces>
        </node>
        ...
    </nodes>
    <links>
        <link id="1 -> 2">
            <from if="10.0.2.0/30" node="10.0.0.1"/>
            <to if="10.0.2.0/30" node="10.0.0.2"/>
        </link>
        ...
    </links>
</topology>
```


TE tools - Opnet SP Guru

- Commercial (www.opnet.com)
- “SP Guru MPLS Traffic Engineering feature automatically defines explicit routes that minimize maximum link utilizations under normal conditions, and secondary routes that will survive link and node failures.”