The Properties and Mathematics of Data Transport Quality

A Brief Introduction to 'Quality' in Data Networks; its Interaction with End User Experience, its Conservation, Propagation, and how it can be Traded, Costed and Managed.

Neil Davies

Predictable Network Solutions Limited
neil.davies@pnsol.com

Ofcom, Riverside House
5th February 2009

Outline

1. Delivering "Quality"
    - Layered Viewpoint
    - "Would you Like Quality with that, Sir?"
    - Relationship with End User Experience
2. Quality Attenuation
    - Fundamental Properties
    - Representation and Measurement
    - Compositional Properties
3. Exploiting the Understanding
    - Applying it to the Application(s)
    - Applying it to the Network(s)
    - Applying it to the Economics

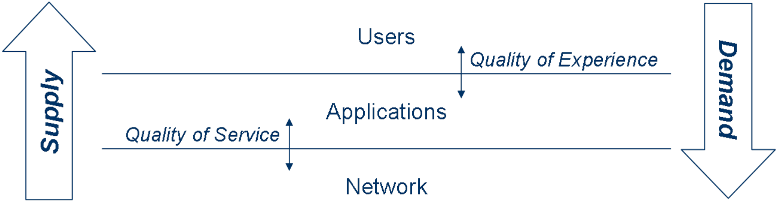
Outline

1. Delivering "Quality"
   - Layered Viewpoint
   - "Would you Like Quality with that, Sir?"
   - Relationship with End User Experience
2. Quality Attenuation
   - Fundamental Properties
   - Representation and Measurement
   - Compositional Properties
3. Exploiting the Understanding
   - Applying it to the Application(s)
   - Applying it to the Network(s)
   - Applying it to the Economics

Delivering Quality

Layered View

For an end-user to achieve a certain quality of experience, an application interacts (with a server or another application) across the network.



For any particular application, the quality the user experiences will depend on how quickly the application can interact (with the remote peer) across the network.

Delivering Quality

Not Just Quantity – Some Frequently Asked Questions:

1. Doesn't it depend on the specific application? Yes and no.
   Badly designed or written applications can make things worse, however the delivered end-to-end quality is now typically dominating the delivered quality of experience.
2. Isn't more bandwidth more quality? No.
   It doesn't matter how much bandwidth you deliver, if the delay is large (or rapidly varying) enough or the loss rate is high enough then the application will fail.
3. So why do people keep on talking about adding bandwidth as the answer? Adding more resources may resolve some issues under limited circumstances. We'll return to this point later.

Any given application's effectiveness depends on end-to-end quality being available in sufficient quantity – no more, no less.

Outline

1. Delivering "Quality"
    - Layered Viewpoint
    - "Would you Like Quality with that, Sir?"
    - Relationship with End User Experience
2. Quality Attenuation
    - Fundamental Properties
    - Representation and Measurement
    - Compositional Properties
3. Exploiting the Understanding
    - Applying it to the Application(s)
    - Applying it to the Network(s)
    - Applying it to the Economics

The 'Just Add Quality' Myth

What has Silence, Cold, Dark and Quality all got in common?

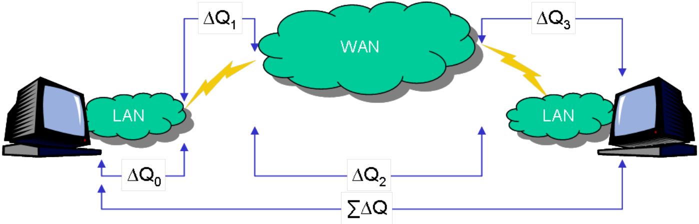
- You can no more 'add quality' to a network than you can 'add silence' to a noisy room.
  Just as silence is the absence of noise, what is colloquially called 'quality' in data networking is really the absence of something.
- Every network element attenuates the quality - introduces delay and (the potential for) loss – every transmission line, switch, router etc.
- People may talk about quality and even desire it, but *quality attenuation* is the physical property we have to work with.

This is a key concept – having introduced quality attenuation we can start re-framing the issues in a coherent framework.

Delivering Quality ≡ Bounding Quality Attenuation

Introduction to Properties of Quality Attenuation (ΔQ)

In data networks, 'quality of service' is achieved when the delivered quality attenuation, over the end-to-end path, is suitably bounded.



We use the concept of quality attenuation so frequently that we refer to it as 'ΔQ' - think of it as the change in quality.

This inevitable ΔQ comes in two forms: *immutable* – fixed by physics, and *mutable* which can be managed and traded.

Outline

1. Delivering "Quality"
   - Layered Viewpoint
   - "Would you Like Quality with that, Sir?"
   - Relationship with End User Experience
2. Quality Attenuation
   - Fundamental Properties
   - Representation and Measurement
   - Compositional Properties
3. Exploiting the Understanding
   - Applying it to the Application(s)
   - Applying it to the Network(s)
   - Applying it to the Economics

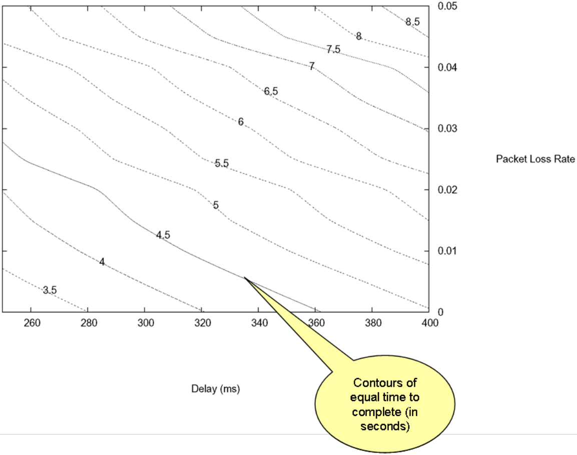
Examples – 1

What Bounded Quality Attenuation Delivers

You want to assure some average performance for typical (10kb) HTTP web page access – What $\lceil \Delta Q \rceil$ should you aspire to deliver?

- What is the dependency on the one-way delay?
- What is the dependency on the loss rate?

The dependency is on both delay **and** *loss, not delay* **or** *loss*



Contours of equal time to complete (in seconds)

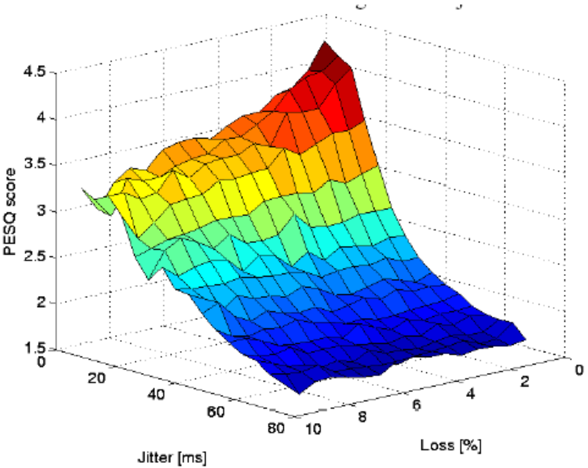
Examples – 2

Applies to Real-time Services As Well

You want to assure some perceived quality for a G.711 VoIP call – what ⌈ΔQ⌉ should you aspire to deliver?

- What is the dependency on the one-way delay?
- What is the dependency on the loss rate?

The dependency is on both delay **and** *loss, not delay* **or** *loss*



Outline

1. Delivering "Quality"
   - Layered Viewpoint
   - "Would you Like Quality with that, Sir?"
   - Relationship with End User Experience
2. Quality Attenuation
   - Fundamental Properties
   - Representation and Measurement
   - Compositional Properties
3. Exploiting the Understanding
   - Applying it to the Application(s)
   - Applying it to the Network(s)
   - Applying it to the Economics

Quality Attenuation

Properties of ΔQ

- It is ***Conserved***
  - it only every increases, and can't be 'destroyed';
    - you can't 'un-delay' packets or 'un-loose' them.
  - hence is monotonically increasing – 'adds', but not by simple arithmetic.
- Manifests itself in two different ways:
  1. ΔQ associated with the data transport for a single user or application instance — an application's viewpoint.
  2. ΔQ associated with a network element (for example a switch/router where multiplexing occurs) applying to all the streams of data that are flowing through that point — a network operations viewpoint.
     - The total ΔQ at that network element is still conserved - but can be 'traded' through differential allocation amongst the individual data streams.

Outline

1. Delivering "Quality"
   - Layered Viewpoint
   - "Would you Like Quality with that, Sir?"
   - Relationship with End User Experience
2. Quality Attenuation
   - Fundamental Properties
   - Representation and Measurement
   - Compositional Properties
3. Exploiting the Understanding
   - Applying it to the Application(s)
   - Applying it to the Network(s)
   - Applying it to the Economics

Representing Quality Attenuation – 1

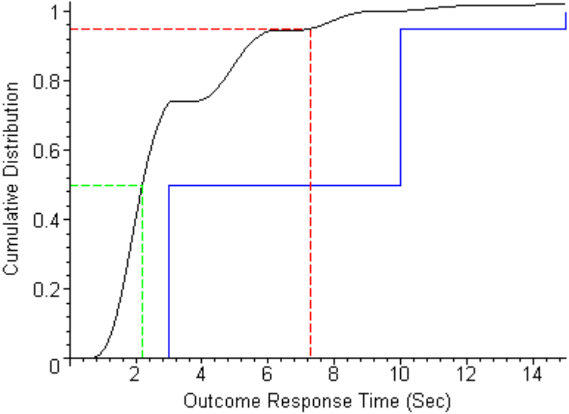
Observing and Predicting Outcomes

From a performance / ΔQ point of view; the interest is in *outcomes*.

If event *A* should lead to *B* occurring, the measure is:

- how frequently *B* actually occurs
- the time interval between *B* and *A*.

We can express both the aspiration (here 50% outcomes occurring within 3s, 95% within 10s - the blue stepped line) and what was delivered (the black line).

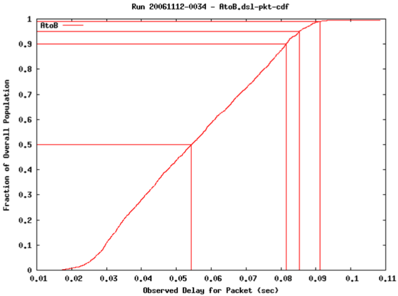


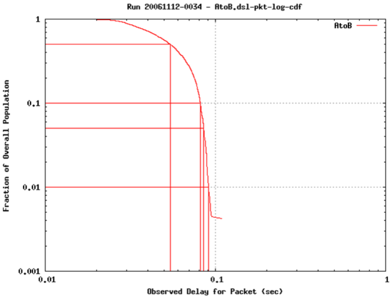
As the delivered curve is always to the left and above the aspiration curve — the aspiration was met and 'quality' delivered.

Representing Quality Attenuation – 2

Focusing on the 'Tail'

It is the tail of distributions that is of most interest[1]





These graphs represent the same outcome: 50% delivered with 54ms; 90% within 82ms; 99% within 91ms; with 0.5% packet loss.

[1] *Maths Note:* This a Cumulative Distribution Function (CDF), technically they are improper CDFs as $\mathcal{P} \not\to 1$ as $t \to \infty$.

Representing Quality Attenuation – 3

Comparing What is Actually Delivered

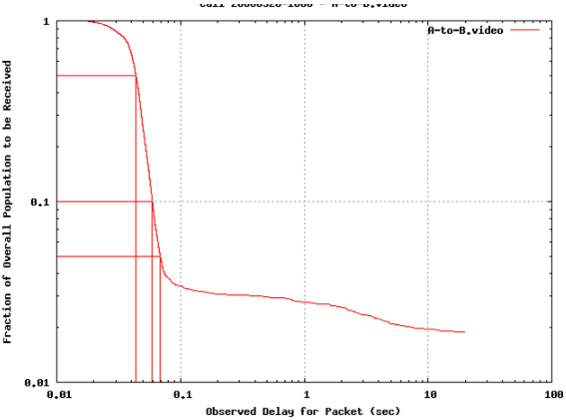
Compare the previous slide's delivered ΔQ:

- *50% within 54ms; 90% within 82ms; 99% within 91ms; ⌈105ms⌉ - 0.5% loss.*

with this graph.

- *50% within 43ms; 90% within 60ms; 95% within 70ms; ⌈2000ms⌉ - 2% loss.*

Same ISP, same application, same two end points, just different time of day.



Quality Attenuation as the Representation Measure

Focusing on the quality attenuation (ΔQ) – especially when the upper bound ($\lceil \Delta Q \rceil$) an application can tolerate is known – is the key.

- Importantly, *All* application requirements can be reduced to this form – it says:

> *"Within this quality attenuation from A to B*
> *deliver to me this (minimum) rate"*

We have a 'budget' ($\lceil \Delta Q \rceil$) to work within!

- How can this budget be divided? How is it allocated across the network elements on the end to end path? What is a reasonable expectation on, for example, the access network?

Outline

1. Delivering "Quality"
   - Layered Viewpoint
   - "Would you Like Quality with that, Sir?"
   - Relationship with End User Experience
2. Quality Attenuation
   - Fundamental Properties
   - Representation and Measurement
   - Compositional Properties
3. Exploiting the Understanding
   - Applying it to the Application(s)
   - Applying it to the Network(s)
   - Applying it to the Economics

Quality Attenuation Budgets

How Aspects of the End to End Path Contribute

The contribution of any network element can be broken down into three components:

- G - dependent on geographical and other fixed factors.
- S - dependent on the packet size and the transmission media.
- V - the variability; dependent on many factors, see below.

1. G is a constant for a given path, incorporates factors such propagation delay and residual error rates for transmission media. It is immutable.
2. S is fixed for a given packet size over a particular path (given that path is fixed) - it captures the delay of processing packets. It is immutable.
3. V is the effects of the rest of the network on this traffic - this is mutable and often highly variable - it is this component that requires management.

These 'sum' (convolve) component-wise for each network element traversed.

Composing ΔQ

If it helps you can think of:

- G as being the time for a packet of zero length to get from A to B (a packet that pays no serialisation/de-serialisation overhead but has to gain access to the transmission medium).
  - ADSL that would be [0–1.5ms][2] (256k) + propagation time; UMTS that would be [0–10ms] + propagation time.
- S being the time to transmit a packet of a given size, this is dependent on packet size and the level-2 networking technology overheads (e.g. quantisation for ATM, frame transmission time in wireless) and incorporates any time that it takes the transmission medium to become available for the next packet/frame (inter-frame gap)
  - This gap is 0 for ADSL and UMTS but a fixed 9.6μs for 10mbps Ethernet.

[2] Uniform distribution between the bounds

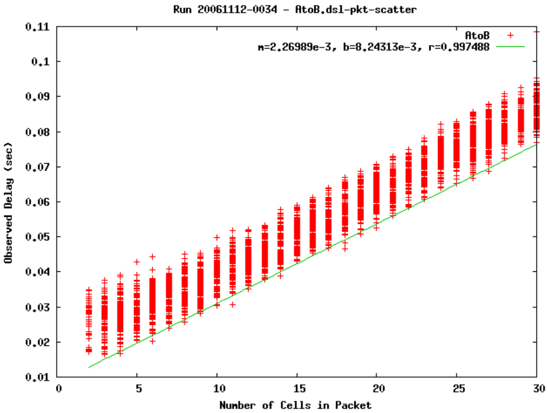
Measuring G, S and V

Beginning to See the Art of the Possible

Here we have taken a sample run and grouped the times by packet size (in this case the number of ATM cells). From this we can deduce:

- G $\approx 8.2ms$
- S $\approx 2.3ms\ cell^{-1}$
- V $\in [0 \cdots 20ms]$

It is the magnitude of *V* that determines the customer experience



Both ends are 256/512k ADSL tails using IP Stream, one in BA6 the other in CT2 going via an ISP based Telehouse North – the Central link was not being used for anything else. There was $\approx 0.5\%$ *packet loss.*

Compositional Properties – Taking Stock

1. Got the tools to measure and analyse where ΔQ is accruing.
    - or alternately divide up a $\lceil \Delta Q \rceil$ budget and allocate to the elements of the network
2. Know that high quality services are feasible (there can be a reasonable bound on ΔQ in access networks)
3. Key to delivering quality is managing and controlling V
    - can't eliminate V; it comes with using statistical multiplexing

How do we 'tame' V?

- Need to look a little more deeply into some other properties of ΔQ.

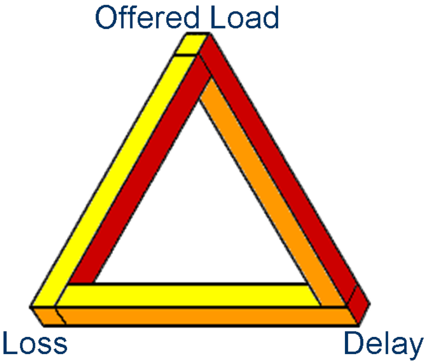
Why ΔQ, not 'Delay and Loss'

Two Degrees of Freedom

Every queue has two degrees of freedom.

- Fix two parameters and you've fixed the third
- Fix one parameter and you establish a relationship between the other two
- Can't choose any arbitrary three values at will

For a fixed load, if you want to reduce loss you have to increase delay; For a fixed loss, as load increases delay must increase; and so on



That 'Question'

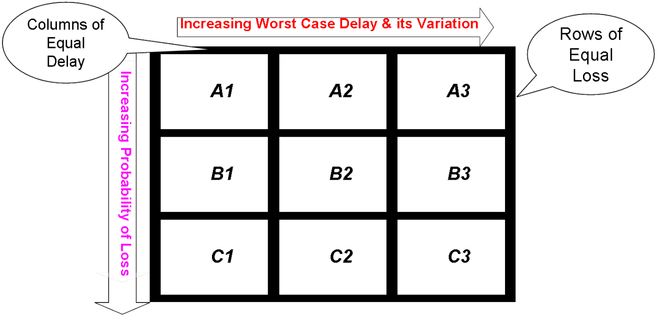
Why do People See Bandwidth as the Answer?

"So why do people keep on talking about adding bandwidth as the answer?"

1. Providers are not managing V, they are taking what 'emerges' from the day-to-day operation of their network.
   - *We'll come to what that means for the consumer and the operator shortly*
2. The V they deliver to their customers is an arbitrary and un-manged relationship between delay and loss
   - 2° of freedom along with the offered load creeping up day by day
3. Their customers complain because their applications are not 'delivering sufficient Quality of Experience'
   - The delay and loss, ΔQ, they are delivering to their customers is too high (over the application's implicit budget)
4. They increase the capacity of links/network to reduce the offered load
   - while the physics allows and they can afford it
5. Because of (1) they return to step (2) and iterate
   - Hence ISP — the Internet (in-)Solvency Problem

2° of Freedom ⇒ Trading Space with Two Dimensions

Trading in Loss and Delay – ΔQ as partial order



Columns of Equal Delay

Increasing Worst Case Delay & its Variation

Rows of Equal Loss

Increasing Probability of Loss

A1	A2	A3
B1	B2	B3
C1	C2	C3

This is how trading within a given ΔQ (or at least the V component) can be visualised. Individual data streams can be given different loss and delay characteristics, so that where contention for resources occur (which is where queues are in network) the resulting ΔQ can be differentially distributed.

For example: traffic in B2 gets lower loss than traffic in C3, but equal delay, lower delay than B3, but equal loss and both lower delay and loss than C3.

Quality Trading in Data Links

More Properly: Quality Attenuation Trading

The properties described above have many interesting consequences on what is possible, or more valuable, what is not possible with data networks.

- One of the more interesting consequences is that any 'pipe' (a path over which data can be delivered within a bounded ΔQ) can carry multiple, differentiated, data transport services even though the 'pipe' itself doesn't support differentiation.
- Alternately, given a multiple data streams the characteristics of the 'pipe' needed can be calculated, so that the collected set of traffic can be carried with that all their individual ΔQ constraints met.

This means that a 'differentiated service' network can be built on top of an existing 'single service' network — if you understand the characteristics and constraints properly.

This offers an incremental (hopefully lower cost) route to delivering differentiated services. Which is useful as differentiated services are essential for the long term economic viability of data networks

Outline

1. Delivering "Quality"
   - Layered Viewpoint
   - "Would you Like Quality with that, Sir?"
   - Relationship with End User Experience
2. Quality Attenuation
   - Fundamental Properties
   - Representation and Measurement
   - Compositional Properties
3. Exploiting the Understanding
   - Applying it to the Application(s)
   - Applying it to the Network(s)
   - Applying it to the Economics

Delivering Quality

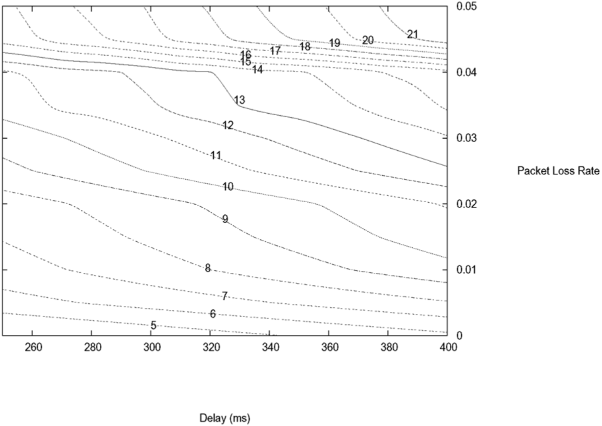
There is No Quality in Averages

Averages are dangerous. People do not remember 'averages' they remember extremes.

- So delivering quality to users is about making bad experiences rare.

This is the graph of the 95% centile of time to complete the same 10Kb HTTP transfer presented earlier.

- This imposes more stringent limitations on ⌈ΔQ⌉



Establishing the Relationship Between QoE and $\lceil\Delta Q\rceil$

So what is the quantity of quality that is needed to achieve some task? There are three basic ways of establishing this:[3]

1. Emulate ΔQ: Connected the parts of the application together through a suitable 'Network Degrader'
    - Expensive, tedious, can be difficult to reproduce faults – however should be part of any validation process
2. Simulate both the application and the network (simulating everything)
    - Expensive, often restricted by computation, supplied libraries (for network protocols) often don't behave the same way as real implementations
3. Analytically. Mathematically model behaviour and ΔQ — solve analytically or numerically
    - Cheaper, used to show feasibility and trends. Can be used to formulate hypotheses to be tested by method (1) or (2).

[3] Note for the unwary: most of the tools out there do not work properly - they will introduce loss and delay, but not in the same way a real network will

An Example

Loading Google's Front Page

This is a DNS look up followed by small HTTP transfer, with allowance for server response times.

- Would have a median/75% centile/95% centile time to complete of 0.73s, 0.76s and 0.81s; given the round trip time was in the range 125ms to 200ms.
- This rises to 2.33s, 2.69s, 3.17s if the round trip range was 125ms to 1000ms

The downstream rates need to support this 'quality' vary from 34.3kbps to 10.8kpbs; the same amount of data over a longer time.

- This shows there can be an advantage (to the provider) in giving 'bad' quality — it reduces the instantaneous offered load — conversely 'good' quality can increase both the peak offered load and its variability.

Outline

1. Delivering "Quality"
   - Layered Viewpoint
   - "Would you Like Quality with that, Sir?"
   - Relationship with End User Experience
2. Quality Attenuation
   - Fundamental Properties
   - Representation and Measurement
   - Compositional Properties
3. Exploiting the Understanding
   - Applying it to the Application(s)
   - Applying it to the Network(s)
   - Applying it to the Economics

Nature of 'The Service' in Current Networks

Specifically Access Networks

Current access networks only offer a single service – the service is not one that they 'specified', it is what 'emerges' during operation. In this service an application's data traffic has:

1. No isolation from the effects of other traffic flowing to/from that end user
2. No isolation from the effects of other traffic flowing to/from other end users (or even ISPs)

Which leads to:

1. People shutting down all their applications and disconnecting other computers so that they can play an interactive game, make a VoIP call or stream some video.
2. Really annoys people as there is nothing they can do about it.

Access network providers do try to do something about (2) — BT uniformly shares bandwidth at their BRAS's, ComCast buckets 'heavy' users into a constrained service class.

What consumers need *is assured bounds on quality attenuation for some portions of their traffic — then the application that is wish to use will deliver what they require.*

UK is Well Positioned

Though More by Accident than Design (or Good Engineering Principles Win Through)

These comments are specifically about IP Stream. It is the only access network with sufficient data about its design and operation in the public domain to be able to draw reliable conclusions.

BT's planning rule for capacity between a BRAS and a DSLAM is that an end user should be able to achieve 2Mbps during the busy period, 90% of the time[4].

- This is the specification of an outcome and, as you will now know, there must be an associated delivered $\lceil \Delta Q \rceil$.
- The equivalent $\lceil \Delta Q \rceil$ corresponds to delivering 97%+ of packets with a low delay variation (15ms to 20ms)

Thus, in the UK, over the national data infrastructure, we already have 'pipes' with sufficiently known, and good, properties into which multiple differentiated services can be multiplexed.

[4] the same planning rules have been proposed for 21CN

Outline

1. Delivering "Quality"
   - Layered Viewpoint
   - "Would you Like Quality with that, Sir?"
   - Relationship with End User Experience
2. Quality Attenuation
   - Fundamental Properties
   - Representation and Measurement
   - Compositional Properties
3. Exploiting the Understanding
   - Applying it to the Application(s)
   - Applying it to the Network(s)
   - Applying it to the Economics

Why Single Service Networks Are Bad News

Why to Remain Tenable Broadband Needs to Move to Multiple Services

1. Citizens, Consumers, Commerce and the Government want to get more out of Broadband.
2. Many of the services people want will require stronger upper bounds on quality attenuation.
   - for example video conferencing or highly interactive applications.
3. But a single service network can, at best, only have a quality attenuation bound — therein lies the problem.
   - No industry can afford to structure its business to deliver all its services at the cost point only a few would be willing to pay.

Having data traffic with differing quality requirements is needed to make optimal use of the infrastructure, with the savings that implies.

Differentiated Service Access Network

So, Who Gets to Decide Which Traffic Gets Treated Which Way?

Simple, the End User.

- Only they know the importance of the applications quality of experience to their requirements.
  - The same application can take on different roles, requiring different bounds on the quality attenuation at different times.
- There should be a differential price (though that may not mean a differential charge) for different qualities
  - This creates the appropriate economic feedback to make a rational market.
- Most important is need for a 'scavenger' style class — one where there are no published bounds on ΔQ, a 'below normal' service
- where the delivered rate can be reduced to a trickle for peak periods — would make a substantial difference to the economics of Broadband delivery
  - The price/charge differential would need to be reasonably high to persuade end users to engage.

Costing Differential Services

Exploiting the Two Dimensional Nature of the Trading Space

The two dimensional nature of the ΔQ trading space has one other interesting property. It can be used to calculate a the cost of delivering of quality, using an opportunity cost argument.

1. At any point (network element) in the network there is some ΔQ, that ΔQ is conserved. Giving 'less' ΔQ to some traffic means that the remaining traffic must experience proportionately more.
   - E.g. more traffic in the $A1$ box means that $B2$ traffic must experience a greater ΔQ in proportion.
2. So the $B2$ traffic is not just reduced by the volume in $A1$, but as some of the ΔQ budget has already been 'consumed', even less volume of traffic can be carried in $B2$ and meet the budget.

Conclusions

- We've come at the 'quality' issue in several different ways — all those 'quality' issues can represented in terms of ΔQ — *Quality Attenuation*.
- ΔQ, its conservation and 2° of freedom is the underlying physical property of statistically multiplexed data networks.
  - Any policy, regulation, service specification, network design, application design, etc, etc has to work within its constraints.
- This is good news, it helps define the 'Art of the Possible', partially by showing what is not possible, and partially by
- ΔQ bringing a quantitative basis to many of the contentious issues that surround data networking today, like
  - How to specify requirements, predict performance, manage large scale networks through creating $\lceil \Delta Q \rceil$ budgets, describe service agreements, and cost proposed services.
- It has been (and being) used to design new network elements, create novel services over existing infrastructure and make distributed computer system safer and more reliable.