

Aims

- Screening
- Capture and retain hard-won collective knowledge - “don’t revisit past mistakes”
- Benchmark results to be **consistent** and **reproducible** - “anyone can repeat tests and confirm results”
- Agnostic - “facilitating decision-making, not taking decisions”
- Open framework - “can add/extend over time”

Non-aims

- Not seeking exhaustive tests. Screening, rather than proving correctness (which is probably impossible in any case).



Putative Framework

- Use unambiguous physical measurements e.g. cwnd, RTT, completion time, packet transmissions/loss/retransmissions over an interval. Avoid sterile “fairness” debate.
- Measure performance over a wide range of bandwidths, RTT’s, queue sizes, number of flows and mix of connection sizes. Topology ?
- Place full data (i.e. for all test cases) plus code in public domain - recognises paper space constraints, facilitates diagnosis of problems/discrepancies (e.g. cwnd time histories invaluable as is actual code used.)
- Propose use of standard TCP as a baseline i.e. always part of any tests. Baseline for evaluation, provides consistency check (e.g. between tests by different groups)..
- Simple tests to start with - uncontroversial, consistent, easily reproducible.



Actions ?

- How about starting with a workshop (hands-on experiments) e.g. as adjunct to PFLDNet ?
 - one outcome might be to document a common subset of measurements and tests (aim is to facilitate comparison rather than being prescriptive).
 - under aegis of IRTF ?

