DC-BENCH: A Benchmarking Suite for Datacenter and HPC Interconnection Networks

IBM ZRL rev.0.89 (Draft under construction)

Editor: Mitch Gusat

Abstract

DC-N# is a collection of synthetic traffic patterns relevant for apps running in datacenter environments which is been defined as a benchmark to quantitatively assess new proposals such as scheduling, routing, load balancing, flow and congestion control schemes**.** Whereas benchmarks help in shaping a field, they also have inherent limitations to be considered when conducting measurements and reporting the results.

Our primary target is –by the means of rigorous and verifiable procedures-- to obtain a small set of (easily understood) figures of merit. Ideally, these should unambigously, consistently and concisely characterize the performance of new schemes, as well as the operation of the tested network in terms of efficiency, stability and fairness.

1. Motivation

So far the networking and congestion control community does not have a benchmark to qualitatively and quantitatively assess the merits of various proposals. Many such proposals show good results under simple congestion cases while leaving open the question of more rigurous evaluations, comparative studies and repetability in other modeling environments.

Derived from our switching work and previous experiences w/ RapidIO and IBA flow control and congestion mgnt., we have defined and tested a selected set of traffic scenarios to test the (i) efficiency, (ii) stability and (iii) fairness of congestion control proposals (and not only). This ongoing work we intend to further expand w/ traces and app-cores as a contribution to IEEE 802 and/or an appropriate forum such as SPEC, TPC.

…

2. Metrics and Measurement Methodology

This section will mainly define the methodology of simulation and benchmarking required to obtain results that can be reported with a high degree of confidence.

Distinguish between congestion types based on network's FC elasticity: single bottleneck in lossy (802, IP -> elastic nets) and saturation tree hotspot (lossless -> stiff net).

…

3. Definition of DC-N# Benchmarking Scope and Procedures

2.1. Network parameters (Table)

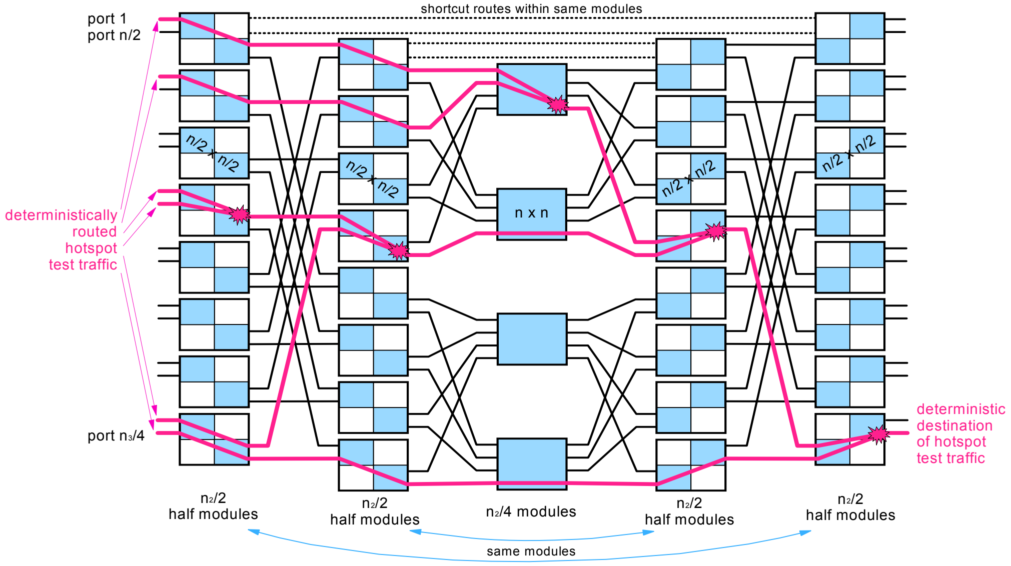
2.1.1. Size:

a) Singe stage: switch degree >= 8 ; link RTT (> 5m); Bw link (>= 1Gbps, raw); MTU (64B-16KB);

b) Multihop: min. size in no. endnodes (e.g. N > 32) and hops (k > 2);

2.1.2. Topology:

1. MIN: Banyan, fat-tree,
2. direct: mesh, torus, cube (dimension)
3. irregular:



2.2. Benchmarking Category and Traffic Pattern Definition

- Outline: Define each test
    - goal / type : stability, Tput, fairness... / static, dynamic
    - matrix equation ∑ λi (i= cold, hot, X-traffic)
    - time phase: start, end
    - metric
    - measurement methodology

1. Category: Defines type and goal of each benchmark along multiple dimensions.

E.g. goal-oriented benchmarks define tests of connectivity, baseline reference, max. performance (Tput, min. Latency), efficiency, stability, fairness.

Along another dimension, the congestion-specific subset contains core, i.e. basic (static, single hotspot) and extensions: composite patterns by superposition of multiple traffic matrixes, possibly dynamic.

2. Spatial distributions. Traffic matrix definition of offered load: Dynamic Superposition

Since only the basic DC-N# benchmarks are static, i.e. time-independent, the general definition also includes a temporal variable *t*. Note, however, that *t* does *not* specify the temporal distribution within a flow (burstiness, message interarrival); instead it defines the benchmark *phase* (aka epoch) during which a certain offered load matrix remains constant. Hence matrix $[\lambda_{ij}]$, possibly multipled by a factor α_k, defines the offered load during benchmarking *phase* $t_{p\text{->}q}$. The corresponding notation $\alpha_k[\boldsymbol{\lambda}_{ij}]:t_{p\text{->}q}$ indicates that from t_p until t_q each ingress source *i* generates messages for each egress destination *j* with rate $\alpha_k\lambda_{ij}$.

Some DC-N# benchmarks are defined as the linear superposition of two or more of the above matrixes, defined during the *each* benchmarking phase. E.g.,

$$\boldsymbol{\Lambda}:t_{p\text{->}q} = \alpha_{\text{k\_hot}}\,[\boldsymbol{\lambda}_{\text{ij\_hot}}]:t_{p\text{->}q} + \alpha_{\text{k\_cold}}[\boldsymbol{\lambda}_{\text{ij\_cold}}]:t_{p\text{->}q} + \ldots + \alpha_{\text{k\_bkgnd}}[\boldsymbol{\lambda}_{\text{ij\_bkgnd}}]:t_{p\text{->}q}$$

$$\boldsymbol{\Lambda}:t_{q\text{->}r} = \beta_{\text{k\_hot}}\,[\boldsymbol{\lambda}_{\text{ij\_hot}}]:t_{q\text{->}r} + \beta_{\text{k\_cold}}[\boldsymbol{\lambda}_{\text{ij\_cold}}]:t_{q\text{->}r} + \ldots + \beta_{\text{k\_bkgnd}}[\boldsymbol{\lambda}_{\text{ij\_bkgnd}}]:t_{q\text{->}r}$$

A pictorial description will also be provided whenever required for better understanding. Note that the resulting traffic matrixes $\boldsymbol{\Lambda}$ will not always generate admissible loads, i.e. when each row and column sums to unity. Instead, the factual purpose of some benchmarks is to precisely define in space and time some of the relevant –to researchers, architects, managers and practitioners--congesting traffic matrixes (inherently inadmissible) .

4. DC-BENCH-N# Suite Outline

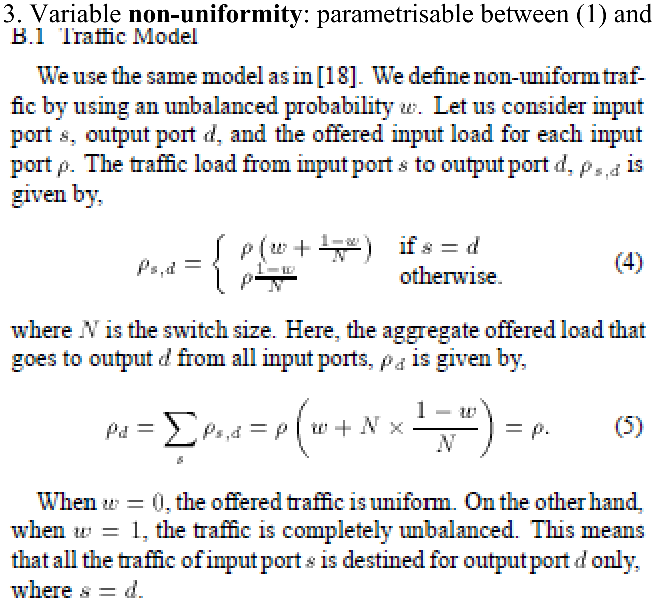
4.1. DC-BENCH-N1: Connectivity and Baseline Performance

Goal: Set the upper bounds {$T_{\text{put Max}}$, $L_{\min}$} of performance

- *A) Non-bursty*

  1. **Contention-free** traffic: (near-neigh. & max. Bbis)
  2. **Uniform**

3. Variable **non-uniformity**: parametrisable between (1) and (2), e.g the w-model from



B.1 Traffic Model

We use the same model as in [18]. We define non-uniform traffic by using an unbalanced probability w. Let us consider input port s, output port d, and the offered input load for each input port ρ. The traffic load from input port s to output port d, $\rho_{s,d}$ is given by,

$$\rho_{s,d} = \begin{cases} \rho\left(w + \frac{1-w}{N}\right) & \text{if } s = d \\ \rho\frac{1-w}{N} & \text{otherwise.} \end{cases} \qquad (4)$$

where N is the switch size. Here, the aggregate offered load that goes to output d from all input ports, ρ_d is given by,

$$\rho_d = \sum_s \rho_{s,d} = \rho\left(w + N \times \frac{1-w}{N}\right) = \rho. \qquad (5)$$

When $w = 0$, the offered traffic is uniform. On the other hand, when $w = 1$, the traffic is completely unbalanced. This means that all the traffic of input port s is destined for output port d only, where $s = d$.

R. Rojas-Cessa, E. Oki, and H. Jonathan Chao, "CIXOB-k: Combined Input-Crosspoint-Output Buffered Packet Switch," *Proc. IEEE Globecom 01,* vol. 4, IEEE Press, 2001, pp. 2654-2660.

Usage: A traffic case will be characterized as N1A (TBD)

- ***B) Bursty*** 10, 30, 50, 70, 90 (s-distribs defined as above)

  1. Contention-free traffic: (near-neigh. & max. B_{bis})
  2. Uniform
  3. Variable non-uniformity

Usage: A traffic case will be characterized as N1B (TBD)

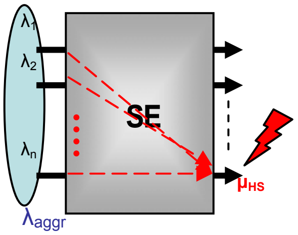
4.2. DC-BENCH-N2: Efficiency (Congestion Kernel Benchmark)

Goal: Measure network behavior and CM's efficiency under a number of congestive loads, statically defined. The method can also be applied to conduct sensitivity analysis during the design of new CM schemes.

A) Single Hotspot

To enforce hotspots we use a model where inputs or sources generate a hotspot at any location within the fabric by offering a predetermined amount of *inadmissible* traffic to the respective location. This traffic load, including its temporal and spatial distribution, is referred to as 'hot'. The general equation

$\mathbf{\Lambda}{:}t_{p\text{->}q} = \alpha_{\text{k\_hot}}\,[\boldsymbol{\lambda}_{\text{ij\_hot}}]{:}t_{p\text{->}q}$, where the matrix $[\boldsymbol{\lambda}_{\text{ij\_hot}}]$ will be specified per each case.



1. **Type** of congestion
   a. *Primary*, aka input-generated (IG)

   b. *Induced*, aka output-generated (OG)

2. Hotspot **severity**: HSV = $\boldsymbol{\lambda}_{\mathbf{aggr}}$ / $\boldsymbol{\mu}_{\mathbf{HS}}$, $\lambda_{\text{aggr}} = \sum \lambda_i$ at hotspotted output, μ_{HS} = service rate of the HS, i.e., the drain rate of the bottleneck link during the congested phase.
   a. *mild*: HSV < 2 (i.e. $\Sigma|\mathbf{\Lambda}{:}t_{p\text{->}q}| < 2\mu_j$)
   b. *moderate*: HSV = 3..5
   c. *severe*: HSV > 10.

3. Hotspot **degree**: HSD is the fan-in of congestive tree at the measured hotspot
   a. '*Small*': HSD <10% of all sources inject hot traffic;
   b. '*Medium*': HSD 20..60% of sources inject hot traffic;
   c. '*Large*': HSD >90% of sources inject hot traffic. (Q: How to show HSD in the traffic matrix?)

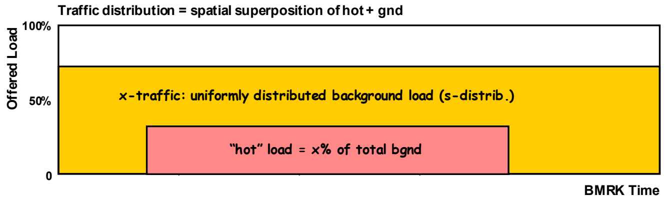
Usage: A traffic case will be characterized as N2A (α_{k_hot}*, type, severity, degree*), e.g. N2 (0.55, IG, HSV=1.1, HSD = 2).

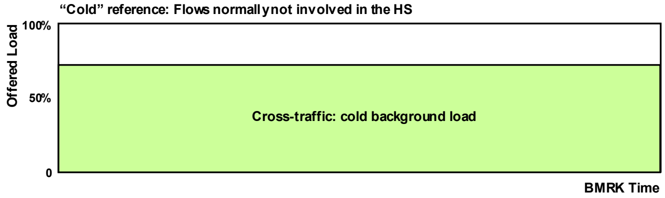
B) Composite Hotspot

Is a superposition of (A) with background cross-traffic of variable load. Thus the traffic profile for this benchmark has two components, i.e. uniformly distributed background traffic and the hot traffic. All input ports generate a given constant $\alpha_{\text{k\_bkgnd}}$ offered load during the simulation/benchmarking time. The test will be repeated for $\alpha_{\text{k\_bkgnd}} = 0.1$ up to the $T_{\text{put Max}}$ determined with N1.A.2.

During a defined congestion period, predetermined hotspot-generating input ports direct a fixed portion of their load to a defined hotspot destination:

$$\mathbf{\Lambda}{:}t_{p\text{->}q} = \alpha_{\text{k\_hot}}\,[\boldsymbol{\lambda}_{\text{ij\_hot}}]{:}t_{p\text{->}q} + \alpha_{\text{k\_bkgnd}}[\boldsymbol{\lambda}_{\text{ij\_bkgnd}}]{:}t_{p\text{->}q}$$


Traffic distribution = spatial superposition of hot + gnd
Offered Load
100%
50%
0
x-traffic: uniformly distributed background load (s-distrib.)
"hot" load = x% of total bgnd
BMRK Time


"Cold" reference: Flows normally not involved in the HS
Offered Load
100%
50%
0
Cross-traffic: cold background load
BMRK Time

Usage: A traffic case will be characterized as N2B (TBD).

4.3. DC-BENCH-N3: Stability

Traffic matrix is dynamically redefined by using either On/Off flows (Markov birth and death process) or *sweeping* patterns.Traffic is injected into the system as packets of fixed MTUs between 64B and 16KB – except for flow control and signalling pkts such as the ACK and ECN packets that are used to measure the impact of shorter packets. In the absence of congestion the inter-packet time is negative-exponentially distributed.

A) Markov-modulated On/Off hotspot traffic. First reason is that the synchronized pattern can expose control loop instabilities and oscillatory behaviour. Equally important is to focus the BMRK on more common traffic scenarios. The On/Off pattern we consider representative for a large class of commercial and HPC applications that use collective and synchro operations – e.g. locks and barriers – or frequently exercise reliable multicast transmissions.

The On/Off is either (1) random or (2) synchronized on multiple of RTT.

Usage: A traffic case will be characterized as N3A (TBD)

B) Sweeping hotspot

A deterministic form of admissible traffic, denoted as *sweeping hotspot.* As the name suggests, this is a traffic pattern whereby all input traffic sources synchronously target one output after another in an arbitrary, but constant, sequence. A round-robin sweeping sequence is shown in Fig. 2. More details in [MMJ paper].

If the k-th output is currently hotspotted, then $\sum_{i=1}^{N} \lambda_{ij}(t) = \begin{cases} N \cdot \lambda_{\max}, \; j = k, \\ 0, \; j \neq k. \end{cases}$

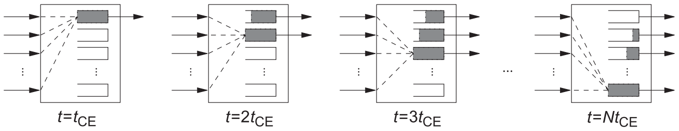


Figure 1. Graphical description of a sweeping hotspot in RR pattern.

Usage: A traffic case will be characterized as N3B(TBD)

4.4. DC-BENCH-N4: Fairness

A) Single HS

1. "Transistor" test: cold 'elephants' matrix + hot 'mice' matrix + cold matrix
2. Near / remote superposition of hot + cold traffic matrixes

B) Multiple HS's

1. Same SRC simultaneously issues both mice and elephants as in A1
2. Same SRC simultaneouslyinjects both near and remote flows

TBC…

ACKs
We're grateful to Cyriel Minkenberg, Wolfgang Denzel, Greg Pfister, Ton Engbersen, Ronald Luijten and many of our ZRL and STG colleagues in US for significant contributions, discussions, drawings and simulations re. flow and congestion control.