HOW MUCH IS ENOUGH?
Rightsizing HPC in Canada
December, 2009

1. Introduction

The utility and need for high-performance computing as a critical component of the research (cyber-) infrastructure of advanced knowledge-based economies is well established. From climate and environmental modeling and prediction to the rapid design and prototyping of new products in industry, HPC is an essential enabling technology. What is much less clear is what level of HPC is appropriate, particularly in an academic environment. Data from the HPC consortia within Compute Canada show clearly that:

- In most cases an equilibrium is established such that academic HPC systems are fully utilized 100% of the time and such that there are a number of queued jobs waiting to start. Unless the demand for HPC exactly equaled the supply, the queues should either shrink to zero or grow indefinitely. This does not happen. The system appears self regulating in that users will not submit additional jobs to the queues if the waiting times become too long but that there is always sufficient work to be done that the queues rarely become empty.
- New systems become fully utilized within a few weeks of becoming available to the community even if the new capacity represents an order of magnitude increase over that previously available.

In this context it is naturally confusing for funding agencies to determine the correct level of support and, equally, it is often difficult for the research community to make a compelling case for any particular level of HPC resources. In some cases, very specific needs can be established but where the HPC resources are intended to support a very wide of research programmes in a shared environment across the whole of Canada, determining the level is difficult. (It is challenging often to determine the HPC needs of a single project and yet Compute Canada must argue for needs aggregated across geography and disciplines.) These notes are intended to provide some data to support a productive discussion.

2. Baseline Research Activity

Given that Compute Canada resources must support such a wide range of research across the whole country, the approach taken here is to look at broad, national, measures of research activity within a number of countries rather than attempting to accumulate the disparate needs of many individual researchers and disciplines. The analysis looks at a number of markers of research activity to try to determine if there are good measures of overall activity that could be used to compare levels of HPC in various countries relative to their research activity and perhaps needs. Clearly some sort of normalization is needed between countries with vastly different activity to make a useful comparison: it is not useful, for example, to compare the absolute activity of Canada either to that of the US or to that of Finland. Data for GDP, number of researchers, percentage of GDP spent on research, number of papers and number of citations is readily available. Figure 1 shows four measures of a country's research activity normalized to a county's GDP. The resultant ratios in each case have been normalized to the mean of the values for the six G8 countries with the highest GDP (thus excluding Russia and

Canada, what will be called the “G6”). This choice affects only the smoothness of the lines in Figure 1 and not the relative values or ranking for any measure. (Russia was excluded because it is anomalous in several respects relative the rest of the G8 as can be seen; the exclusion of Canada changes very little as it is already close to the mean.) Also shown, for interest, is the ratio of population to GDP normalized in the same way. Note that the legend is in order of decreasing GDP. The figure highlights with thick, solid lines the 12 countries comprising the union of the G8 and G10, our natural comparators/competitors. China and India are shown with thin solid lines, the remaining countries – not a complete list by any means and often with incomplete data – are shown with dotted lines. These latter countries are included because they have had some HPC presence over the last few years. It is notable that most of the developed/industrialized countries lie close to the locus mapped out by the G8/G10; countries falling far from the mean for any measure are named: none of these excursions seems particularly surprising. Finally, the two horizontal orange lines indicate a factor of two above and below the mean for each category. The intent of presenting the data in this way (which is not claimed to be novel) is to establish that a wide range of measures give a broadly consistent measure of research activity relative to GDP (and as will shortly be seen show a dispersion that is often significantly less than that for the equivalent HPC data). There are many caveats, of course, but the assertion is that, overall, GDP is a good measure (to within a factor of $2^{\pm 1}$) of research activity in the industrialized countries and that Canada is close to the mean. (In fact, as is well known, Canada does relatively well in terms of research outputs such as publications and citations despite having somewhat lower funding levels.) We will now consider how the provision of HPC varies with GDP amongst countries.

3. Measuring HPC

The fundamental problem with comparing investments in HPC across countries is the virtual impossibility of determining the funds that have been expended on HPC in each jurisdiction. In order to arrive at a useful measure, we turn to the Top500 list and measure by country the simple statistic of aggregate capacity in peak GFlops of all systems on the list in a given year. There are, of course, a huge number of biases that may be present in this data or that will be introduced by using this statistic including varied reporting, types of system, vendor discount levels in various countries, non-academic systems (the New Zealand value on the list, for example, is currently dominated by film industry systems), etc., etc. Nonetheless, unless one can argue convincingly that there are dramatic variations in any of these factors among the large-economy, industrialized countries, it seems a reasonable assertion that this statistic will provide a useful relative measure of HPC investment available to the research community in a particular country. It is certainly direct and avoids trying to understand which portion of investment announcements in various countries will actually end up putting resources on the ground for the benefit of end users. (One obvious bias that may greatly reduce the reliability of the statistic for smaller countries is that the Top500 list cuts off at an absolute threshold and so these countries may have very few systems even appearing on the list and so it is possible that their aggregate capacity could be significantly misrepresented. In defense of the threshold, one may argue that we are discussing HPC rather than just computing and so imposing a minimum performance limit for a system to be counted may not be an unreasonable way to determine a country’s access to *competitive* HPC power. Whether or not the specific level imposed by the 500th highest system is appropriate is questionable, of course, but is, unfortunately, moot for the present analysis. We chose the peak GFlops numbers reported as perhaps having a closer relationship to dollars invested but this also is certainly debatable.

Fig 2 shows the aggregate GFlops for countries normalized by GDP (and displayed relative to the mean of this value for the "G6" as before) since 1993 with data taken from the November Top500 lists. (Note that the GDP for 2008 is used: most of the industrialized countries have broadly similar economic growth rates, of a few percent per year, and the discrepancy that this will introduce over a few years will be small relative to the factors of two that we are focusing on and relative to the approximately 85% yearly changes in HPC performance that are seen on the list.) The countries included on this plot are those that lie within the fiducial developed/industrialized band in Figure 1 and thus represent the cohort that have similar research inputs and outputs and amongst which it seems reasonable to compare investments in, and commitment to, HPC

Figure 2 demonstrates that the US makes consistently higher investment in HPC than all other large economies. This perhaps argues that the US should be excluded when calculating the normalizing mean. The recent report by the Council of Canadian Academies on the productivity gap between Canada and the US, however, ascribes a significant part of this gap to the lack of uptake of Information Technologies in Canada. HPC is surely a narrow segment of IT but, given the significant uptake of HPC by US business and the return on these investments that are reported[†], it seems unwise to discount its impact overall. For these reasons, it does not seem particularly valuable to mitigate Canada's relatively poor performance by artificially suppressing the mean.

The heavy red line for Canada also shows a projection of the HPC:GDP statistic over the next 4 years. This extrapolation is made by assuming the normal growth of the aggregate capability of the Top500 list over all (which has been remarkably predictable at an average rate of 1.85 times per year since the list's inception) and by including all systems funded by the NPF, but not yet installed, in their planned year of installation. For these unspent dollars, it is assumed that a given dollar buys a fixed position on the Top500 over time. (For this latter calculation a rate of exchange of 15TFlops/M$ is assumed for 2009.) The arrow roughly at Spring 2013 marks the very earliest that any new system could be expected to be available to the community as a result of a call for NPF2 expected Fall 2010. The years in which funding from the several CFI rounds result in Top500 systems are shown. Finally, the heavy dashed line shows the five-year centrally-averaged values for Canada.

Clearly there is a great deal more scatter in this diagram relative to the fiducial $2^{\pm 1}$–factor band even for the G8/G10: the importance of HPC to the research ecosystem is not equally recognized by different countries. Furthermore, without several new systems coming online every year, the rapid depreciation of HPC technology will inevitably lead to large year-to-year variability and this is likely to be worse for countries with smaller GDPs and smaller installed HPC base. Despite the clutter, a number of useful/interesting features are evident:

- In Canada relatively few university-based (academic) systems made the Top500 list before 2000; the aggregate performance was dominated during this period by systems at Environment Canada (then AES). After this time, the list was dominated by academic systems as discussed below.
- The impact of CFI funding is clear in the smoothed data beginning with CFI Infrastructure round 2 but becoming entrenched with CFI IF round 3.
- Only a few countries, notably the US, UK, Germany and Japan, are able to maintain relatively

† Council on Competitiveness study: http://www.compete.org/news/entry/525/council-on-competitiveness-idc-release-study-on-hpc-and-innovation/

stable (or at least smooth) aggregate HPC capability.
- Clearly not all G10/G8 countries invest equally (relative to GDP) in HPC.

Overall, what is noteworthy is that whilst the latest infusion of funding has moved Canada significantly closer to the mean of industrialized countries, delays in NPF2 funding quickly translate into a significant competitive disadvantage.

In order to smooth some of the scatter evident in Figure 2 and to include a wider cross section of countries investing in HPC, including several countries that would not normally be classed as developed/industrialized, the HPC:GDP ratios (normalized in the usual way) are averaged for the five years from 2005—2009 and plotted as a ranked list in Figure 3. The figure shows the 41 nations that have any presence on the Top500 list on the past 5 years: Canada is 24th on this list. Only one country with an economy bigger than Canada's falls below it on the list (Italy), whereas 16 countries with smaller economies are higher on the list. Only one G8 country and one G10 country are below Canada: Italy and Belgium respectively. (Note that this chart does not include any projections for investments that are committed. Australia for example has just announced that it will expend A$50M on a petascale machine that it is estimated will push it to a position just below Germany in terms of the HPC/GDP ratio.)

4. How much HPC is needed?

It is interesting to try to estimate the annual investment that would be necessary to bring Canadian levels of HPC (relative to GDP) in line with the mean shown in Figure 2. This is challenging since, as noted earlier, the aggregate Top500 scores include contributions from various host organizations several of which may be inaccessible to the academic research community. A compounding problem is that different countries have widely varying mechanisms for funding and providing access to HPC infrastructure. In Canada almost all of the HPC available to academic research is funded through CFI and the provinces at universities and is relatively easy to track. In other countries, significant HPC infrastructure may also be available to the equivalent of Canadian post-secondary-based research in major research institutes or national labs. Further, the reporting of systems on the list is not uniform: even within Canada some systems funded under the NPF are listed as "academic", some as "research". The latter category is usually available to the academic community but not always.

Over the last five years, over all countries, approximately 60% of the aggregate performance on the Top500 resides in the "academic" and "research" categories. The remainder is "industry", 30%, "classified", "vendor" and "others" (together 10%). For the most recent list, a brief examination of the "research" sites worldwide suggests that less than 6% of the capacity is in systems that are not available to serve the equivalent research mandate of the systems funded in Canada by CFI and the provinces. These inaccessible systems are primarily in meteorological and military research centres. The US tends to distort all statistics and the presence of a significant capacity "behind the fence" at some national labs may reduce the fraction that is truly open. The countervailing argument is that much of this capacity is available for certain periods to "outsiders" and the type of research done "inside" often closely aligns with that done "outside". Indeed, there is significant exchange of people, ideas and techniques that benefits the whole research endeavour in the US. A conservative estimate might be to assume that at least 50% of the aggregate Top500 capacity is available to the academic community. In Canada, going back over the last five November lists the fraction has been: 83%, 70%, 44%, 89% and 57%, an average of just under 70%. A few other examples tend to support fractions in the 60-70% range: the five-year average in Switzerland is 66%, in the Netherlands 79% and in Finland 60%. In New Zealand the five-year average is 20% and this reflects the committed and concentrated investment in the film industry that completely dominates the HPC landscape in this very small economy (9% of Canada's GDP in 2008). Given the relative consistency (across our group of comparator nations) of the fraction of the installed HPC base that is available to academic research, it is probably safe simply to assume that a fractional increase in funding for academic research (from CFI and the provinces) would lead eventually to the other 30-40% being supplied by the other sectors (i.e., primarily industry and government). The alternative, to assume that CFI and the provinces must make up the full deficit that Canada exhibits overall in the Top500 is probably unreasonable.

On average, after 2002, the Canadian level of HPC relative to GDP has been 41% of the mean. The investment by CFI over this period – in equipment alone – has been very close to $15M/yr. (This level would have been maintained for the NPF funding if spread over 4 years; it now appears as though it will be spread over approximately 5 years, corresponding to $12M/yr.) A funding level of $30M/yr for equipment (CFI + provinces) will be assumed as the baseline that has been relatively stable since 2002.

To bring the Canadian HPC/GDP ratio to the mean for the industrialized countries would thus suggest an investment by CFI, in HPC *equipment*, of $15M/0.41 = $36.5M/yr or $180M over 5 years or a total investment by CFI and the provinces of $360M over 5 years ($72M/yr). This would place Canada between the UK and Germany in terms of the HPC to GDP ratio. (It is worth stressing that this would not mean that we would have nearly as much HPC in absolute terms as the UK since our GDP is less – in fact we would have around 40% as much – but the availability per GDP$ – or researcher – would be much closer to par.) This figure can be compared with the equipment investment recommend in the Long-Range Plan for HPC which was projected to be $50M/yr in 2010. This level of investment would bring Canada to 62% of the mean and place us between Taiwan and Saudi Arabia in Figure 3. Conversely, an investment of $12M/yr – the level Canada will have fallen to when NPF2-funded equipment becomes available – will drop the ratio to 0.3, leaving us just ahead of Poland.

5. Sources

HPC: Top500 www.top500.org (November lists. Hong Kong added to China; Puerto Rico added to US)

Population by country; GDP (IMF US$): Wikipedia www.en.wikipedia.org

R&D expenditures; Researchers per population: Unesco www.uis.unesco.org (data most recent at or before 2005)

Papers, Citations: Thomson InCites http://in-cites.com/countries/2007allfields.htm

Innovation and Business Strategy: Why Canada Falls Short: Council of Canadian Academies http://www.scienceadvice.ca/documents/%282009-06-11%29%20Innovation%20Report.pdf

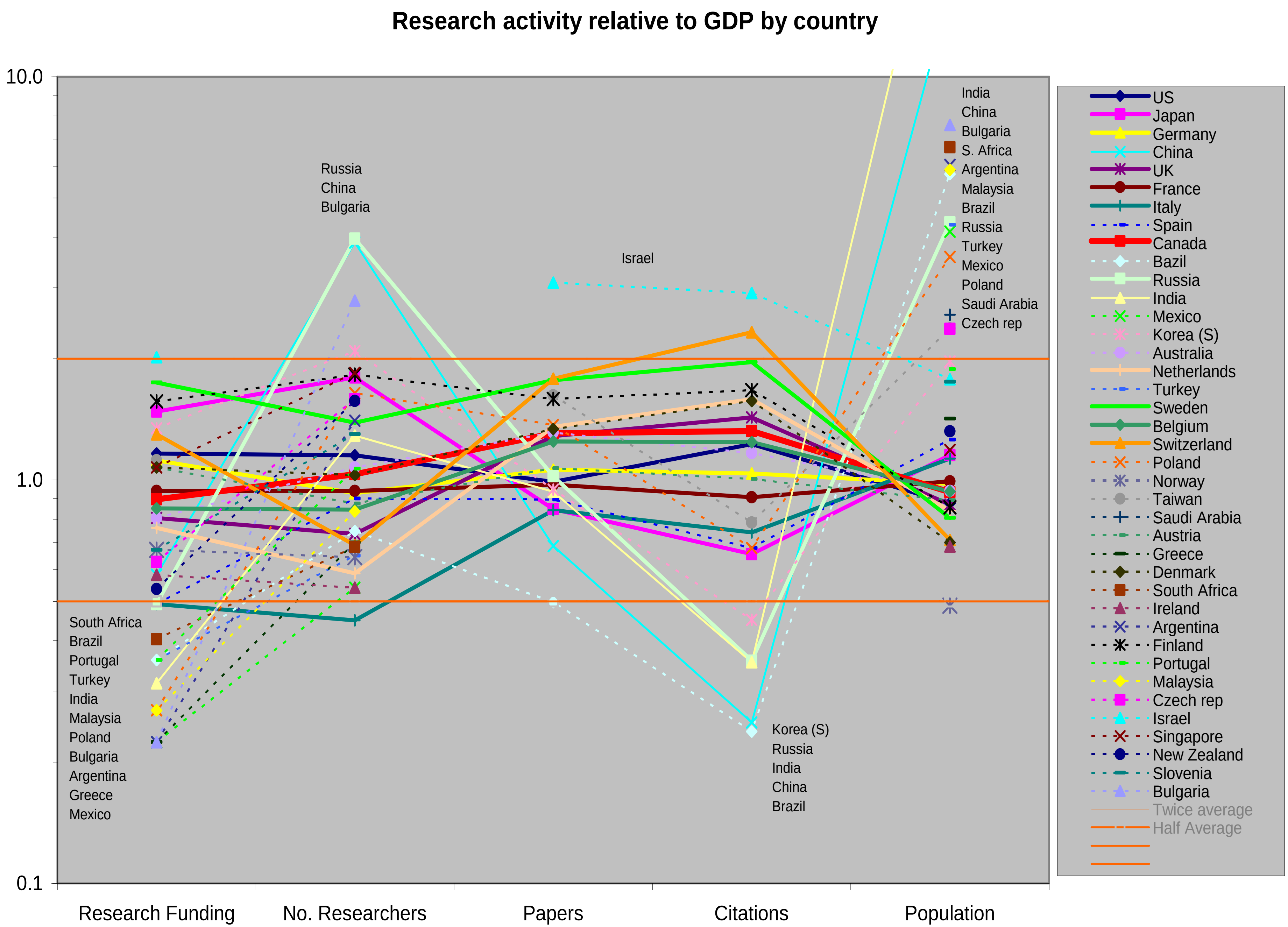

Figure 1 – *see text*

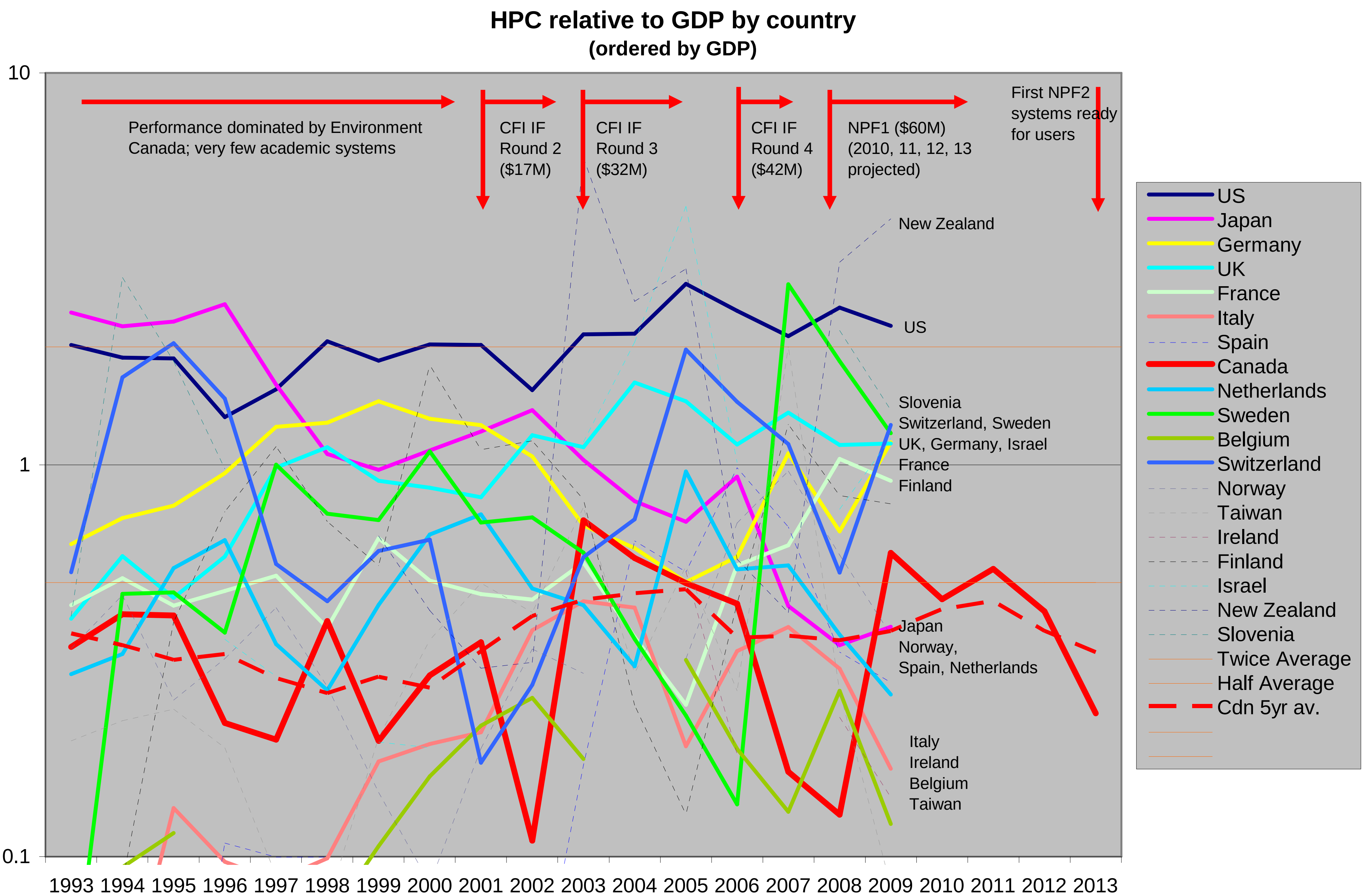

Figure 2 – *see text*

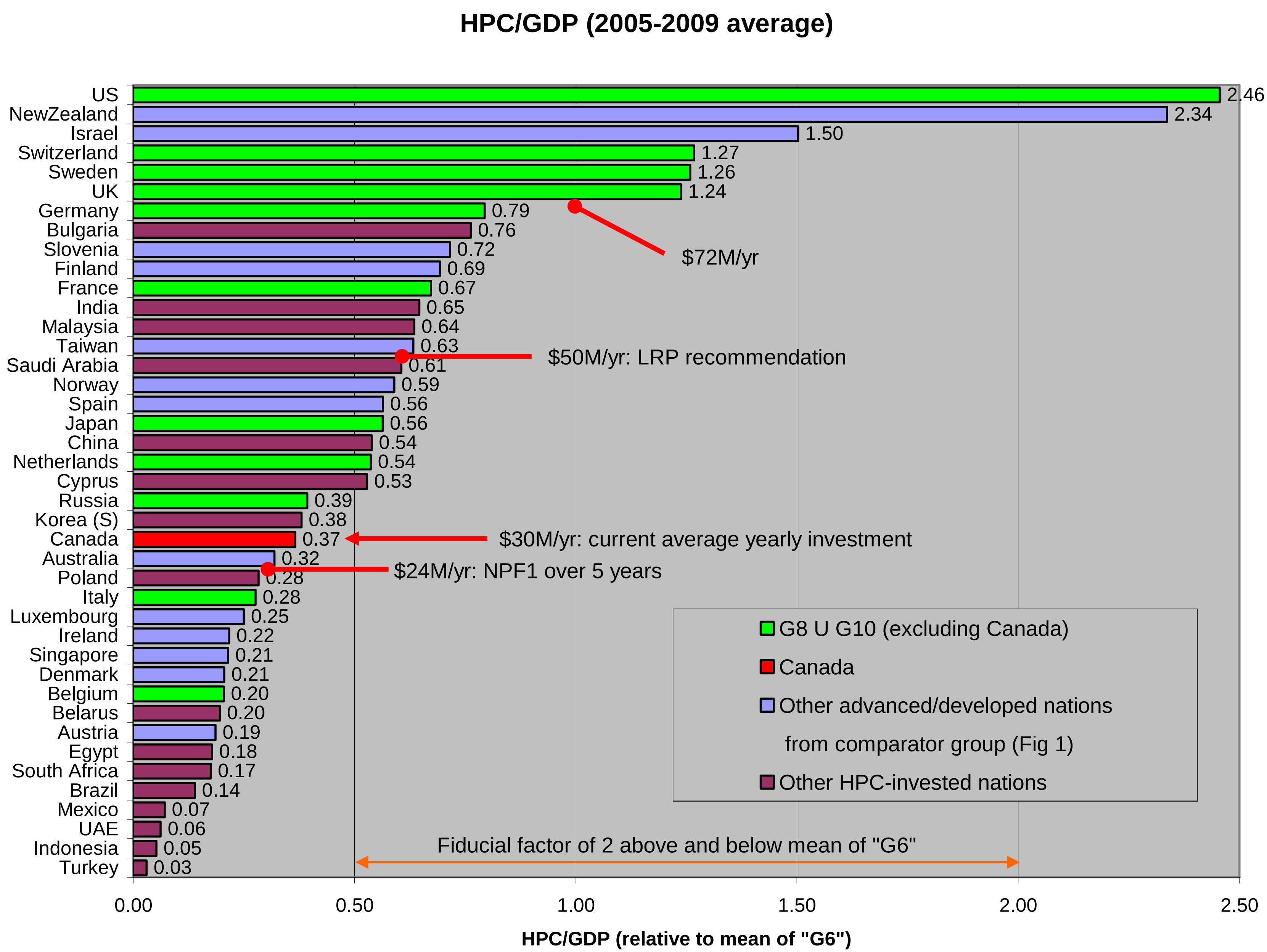

Figure 3 – *see text*