BOSTON
Servers | Storage | Solutions

SUPERMICRO

CFD ARTICLE

KNOW YOUR LIMITS

FORMULA 1'S CFD RESTRICTION REGIME HAS BEEN SHAKEN UP BIG TIME AS THE FIA LOOKS TO CUT THE COSTS OF AERODYNAMIC DEVELOPMENT.

INTRODUCTION

This year Formula 1's CFD restriction regime has been shaken up big time as the FIA looks to cut the costs of aerodynamic development. But has it worked, and howhas it changed both the tools and the process? Racecar investigates

By GEMMA HATTON

Boston worked together with UniFi and CE to benchmark the performance of new CFD technologies in accordance with the 2018 regulations to see if F1 teams would be forced to upgrade their CFD capability.

In January 2018, the FIA introduced the latest evolution of aerodynamic testing restrictions for Formula 1, and with them came the biggest change in CFD restrictions since they were first introduced back in 2009. Racecar went behind the scenes with HPC specialist, Boston Ltd, to discover the impact of these changes and how Formula 1teams have not only benchmarked new solutions, but also upgraded their CFD supercomputers.

But to put these latest changes into context we need to understand the history of the restrictions, both for CFD and the wind tunnel. In 2008, aerodynamic testing was at its peak. BMW Sauber, Honda, Williams & Toyota had all invested huge sums of money in new state of the art full size wind tunnels, each costing tens of millions of pounds. All the top teams were operating in two wind tunnels simultaneously, while Toyota was not only using two wind tunnels 24/7, but each of these was full size.

However, the vast majority of this wind tunnel testing utilised scale models, & over the years the scale of these models increased from 40 per cent to 50 per cent & then 60 per cent.

Operating two wind tunnels full time allowed these teams to complete around 500 wind tunnel simulations per week, with each simulation incorporating approximately 20 different car attitudes. Full size wind tunnel testing was commonplace, with teams either using their own facility or a customer facility such as Windshear in the USA.

IT QUICKLY BECAME CLEAR THAT SOMETHING HAD TO BE DONE TO CURB THE GROWTH OF AERODYNAMIC TESTING IN F1, AND ITS ASSOCIATED COSTS

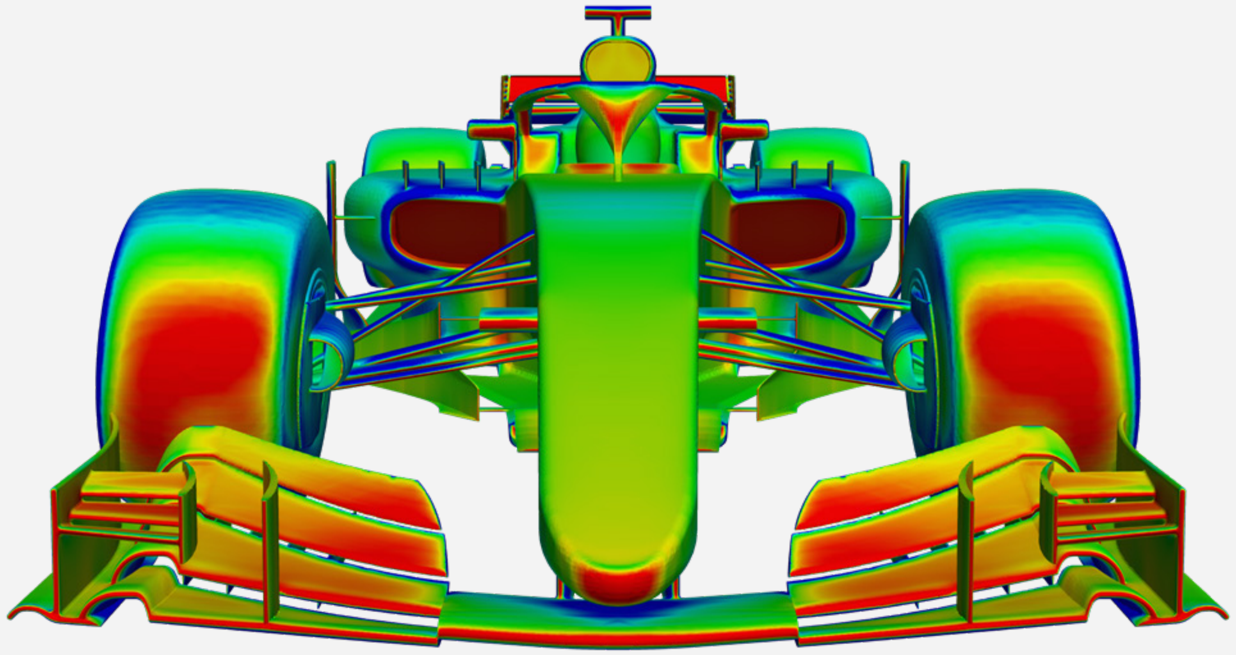
In 2008 teams were already using CFD routinely as part of the aerodynamic development process, and as the software and correlation improved while hardware costs reduced, teams began to use it more, integrating it further into the design cycle. At that time, BMW Sauber was leading the way in CFD hardware with the Albert 3 supercomputer and over 4000 Intel cores, but other leading teams were not far behind. It quickly became clear that something had to be done to curb the growth of aerodynamic testing in Formula 1, and its associated costs. The first step came into force in January 2009 as part of the FOTA Resource Restriction Agreement (RRA).

This controlled the aerodynamic resources the Formula 1 teams could deploy via restrictions on the wind tunnel 'wind on time' (WON) and the CFD compute capacity, measured in TeraFLOPS (TFLOPS). Wind on time was simply a measure of the amount of time the fan was turned on in the wind tunnel with the wind speed in the test section above 15m/s. For CFD, TFLOPS was eff ectively the number of floating point operations completed within the designated eight week Aerodynamic Testing Period (ATP) and was defined by the following equation:

$$TotFLOPs = \left(\frac{\text{MDPPC x CCF x MCU x NSS}}{\text{604,800 x 8 x 1,000}} \right)$$

WHERE:

TotFLOPs =	Total number of TeraFLOPs used per second
MFPPC =	Peak double precision fl oating point operations per cycle per core of the processing unit
CCF =	Peak processing unit clock frequency in GigaHertz
NCU =	Number of processing unit cores used for the run
NSS =	Number of solver wall clock seconds elapsed during the run



Between 2009 and 2017 the regulations evolved and generally served to reduce the aerodynamic resources available to the Formula 1 teams, particularly in the wind tunnel. This was done through introducing a 'limit line' which is defined by the following equation:

$$WT \leq WT\_limit \left(1 - \frac{CFD}{CFD\_limit} \right)$$

WHERE:

WT =	Wind on time
WT_limit =	25 hours
CFD =	TeraFLOPs usage
CFD_limit =	25 TeraFLOPs

Therefore, the amount of time a team chose to run its CFD directly dictated how much time it could utilise the wind tunnel. Equally, if a team could complete its maximum allocation of wind tunnel runs using less wind on time then it would have more capacity for CFD simulations.

WORKING AREA

Looking at the WT_limit and CFD_limit data from the last few years, Figure 1 can be created. Essentially, by plotting the maximum of each of these limits, you can establish the 'working area' that the teams could operate in.

For example, in 2013, when the maximum WT_limit was 60 hours and the maximum CFD_limit was 40 TerFLOPs, the team could operate anywhere within the green shaded area. In 2014, the limits were 30 hours WT and 30 CFD TeraFLOPs, illustrated by the red shaded area, whilst 2015 was limited to 25 hours WT and 25 CFD TeraFLOPs, represented by the blue shaded area, which remained the same until 2018. Since 2013, you can see that overall testing has dramatically reduced, but particularly for the wind tunnel.

For example, let's assume that CFD capacity allows a maximum of 12.5 TeraFLOPs. Using the equation with the 2013 limits results in 41.3 hours of wind on time, as shown by the green square. In 2015, however, 12.5 TeraFLOPs would only give you 12.5 hours in the wind tunnel (blue square) – that's 70 per cent less than 2013. The exact balance between CFD and wind tunnel resources varies from team to team, and sometimes from year to year, depending on the strategic approach and technology advances adopted by each team.

Of course, every restriction that is introduced simply triggers the teams to exploit the loopholes and optimise their designs & working practices to maximise their performance from the regulations. For the TFLOPS CFD restrictions, this became an arms race as teams pushed to develop their supercomputers to run the most CFD simulations per given TFLOP allowance.

This led teams to operate CFD hardware in ways which were quite diff erent from the wider industry, with a clear focus on regulatory efficiency rather than financial efficiency. For example, the TFLOPS calculation naturally includes a chipclock speed term which is reported either as the maximum turbo clock frequency stated on the CPU specifi cation (if the turbo mode is used), or the base clock frequency if the turbo mode is not used.

Teams quickly established that the turbo mode was not an effi cient way to run CFD simulations, in terms of the number of CFD simulations completed per TFLOP. This was also true for many higher clock speed chips. Effectively, running supercomputers with slow clock speed was giving teams more efficiency under the regulations but with the obvious penalty in terms of CFD simulation turnaround time. Therefore, teams then had to balance the speed with which they receive their CFD results against the total number of CFD simulations they were able to complete within the regulatory framework. This is quite different to the wider CFD industry, where the turbo mode was 'free' performance and quicker clock speeds were performance gains if your main criteria was fi nancial effi ciency, and so the divide between the two environments was underway from 2009 onwards.

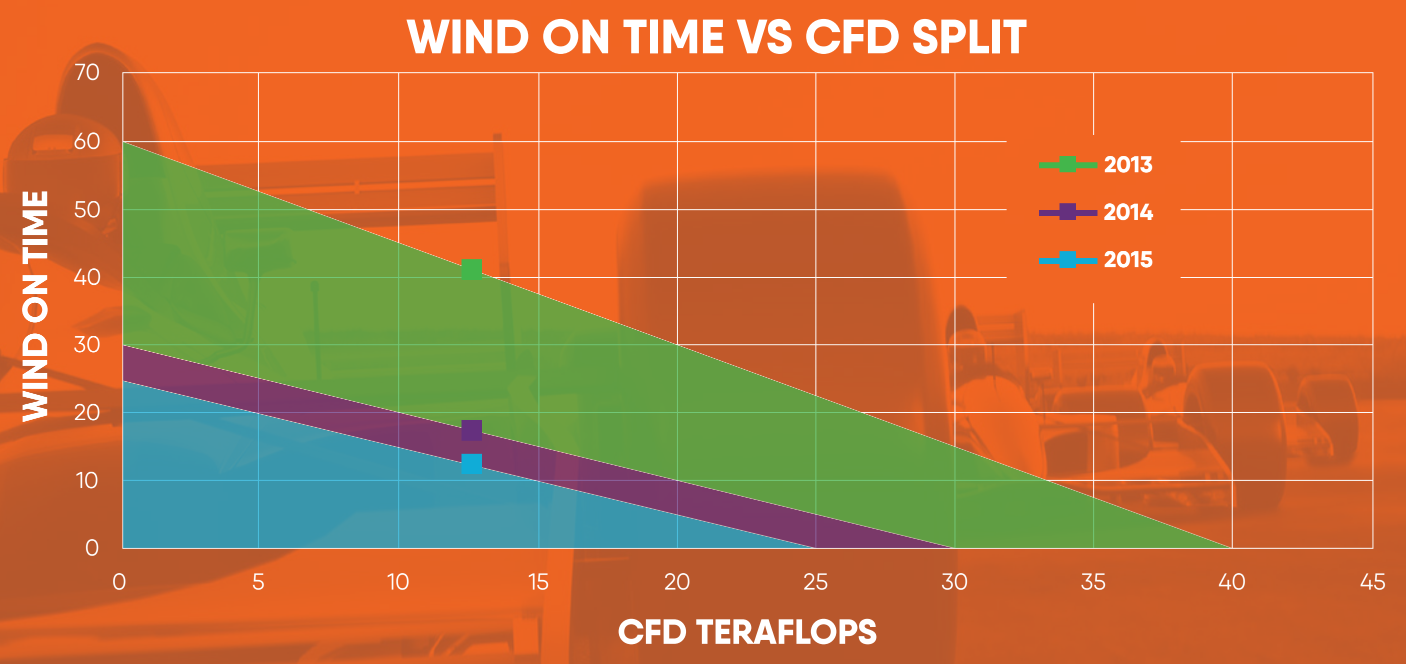
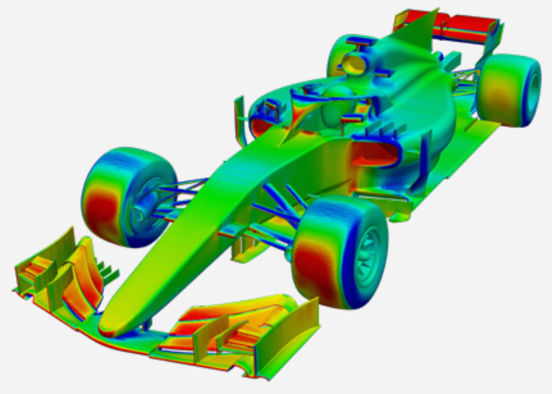


Figure 1: The FIA has restricted aerodynamic testing over recent years for both CFD and the wind tunnel, but particularly the latter. This graph shows the 'working area' that the teams have been able to operate in. Assuming a maximum CFD capacity of 12.5 TeraFLOPS you can see that wind on time has dropped by 70 per cent between 2013 and 2015

CORES & EFFECT

Core under-population also became common place in Formula 1 as it delivered further regulatory effi ciency gains for the teams. It was efficient for the FIA TFLOPS regulation, but it was very inefficient fi nancially, with as much as half of the purchased HPC compute cores being left idle. The biggest issue came when one of the teams developed the Fangio chip in collaboration with AMD, a chip specifically designed to optimise the balance between CFD case turnaround time and throughput which gave that team a huge initial advantage. This exploited the fact that the modern HPC chips were then rated at eight double precision flops/ cycle but commercial CFD codes were only capable of delivering approximately one dp flop/cycle.

The Fangio chip was designed to operate at two dp flops/ cycle giving a big efficiency improvement in MFPPC. Following lobbying from various teams, the FIA agreed to consider the rival Intel chips (Sandybridge and Ivybridge) as four dp flops/ cycle for the purposes of the regulations rather than their rated eight dp flops/cycle. By 2012 AMD had been persuaded by many teams to produce a second limited run of Fangio chips, allowing more of the grid to upgrade their supercomputers to this specification, with most of the remaining teams running an Intel Ivybridge system. With the FIA unwilling to extend the flops/cycle exemption to more modern Intel chips, such as the V3 Haswell CPUs which were rated at 16 dp fl ops/cycle, and AMD not producing any more Fangio chips, the teams were now locked into these older systems purely by virtue of the regulations. Newer chips were simply not viable because of their high fl ops/cycle rating.



Fully correlated and complex CFD models, such as this by Simscale, are becoming an ever-increasing asset to F1 teams, with some full car models now exceeding one billion cells

EVERY RESTRICTION SIMPLY TRIGGERS TEAMS TO EXPLOIT THE LOOPHOLES, ANDOPTIMISE DESIGNS AND WORKING PRACTICES TO MAXIMISE PERFORMANCE.

These older systems were coming to the end of their life and were no longer supported by Intel or AMD. Clearly the FIA had to do something, and the target was to introduce a new regulation which aligned the Formula 1 aero departments more closely with the wider CFD industry as well as allowing teams to upgrade to more modern, supported technology. This resulted in the 2018 CFD restrictions and a move from TFLOPS to Mega Allocation Unit hours (MAUh) as defined by the following equation:

$$AUh = \left(\frac{NCU \times NSS \times CCF}{3600} \right)$$

$$MAUh = AUh \times 1{,}000{,}000$$

WHERE:

AUh = Allocation unit hour

NCU = Number of processing unit cores

NSS = Number of solver wall clock seconds elapsed during the run

CCF = Peak processing unit clock frequency in GigaHertz

Effectively this a very similar measure to TFLOPS but without the reliance on fl ops per cycle, hence removing the barrier to upgrading to newer, better supported, technology. The FIA commissioned an independent study to be carried out in order to set the regulation limit with the intention of giving parity between the old regulations and the new ones. The link to WON was retained and a parallel regulation was introduced with the aim of allowing teams to continue using their old systems if they wished, without too large a performance penalty – at least that was the intention.

Boston Ltd has been specialising in high performance computing (HPC) in a wide range of sectors for over 25 years. In 2017 it formed a new partnership with Tim Milne of UniFi Engineering Services Ltd (UniFi) and Dr Lee Axon of Computational Engineering Ltd (CE). Milne and Axon have extensive Formula 1 experience, most recently at Manor F1 where they were head of aerodynamics and head of CFD correlation respectively.

This group combined Boston's extensive HPC technical knowledge with UniFi's and CE's F1 aerodynamics and CFD experience to provide the F1 teams with a comprehensive benchmarking of the new AMD EPYC and Intel Skylake Platforms. They were able to use all the main F1 CFD codes with models aligned to F1 methodologies and HPC hardware set-ups to extract the maximum possible performance from the new regulatory environment.

NODE TO JOY

The project began in August 2017, by which time Boston Ltd was one of the first companies worldwide to have invested in its own eight node dual socket AMD EPYC system based on the EPYC7601 32 core chips and a similar eight node system based on the Intel Skylake 8176 Platinum 28 core chip. The group also had access to a smaller four node Intel Ivybridge HPC which was used to provide a baseline of the performance gains that teams could achieve by upgrading from their older systems to the new hardware.

This allowed Boston to benchmark its own internal CFD model across a range of CFD codes with a wide variety of hardware set-ups. The systems were all set up with the very latest in networking fabric, up to date operating systems and storage solutions, ensuring that the results obtained would be aligned to the expectations of the F1 teams.

HOT CHIPS

Following the benchmarking of the older Ivybridge system, a number of options within

A typical HPC cluster from Boston. With each new generation of compute chip delivering up to 20 per cent effi ciency improvement the increased capacity of modern CFD clusters means that teams can now have an extra 200 runs, as opposed to 20 back in 2009

the AMD EPYC range as well as the Skylake 8176 chip were evaluated as single node tests to gain an initial assessment of the various different chips available in each family, as well as some insight into the time/iteration performance benefi ts of diff erent options such as the turbo mode.

This also ensured that a clear understanding of the raw performance of the compute chip was gained and that the results were not clouded by any networking issues which could be useful later in the process when trying to understand the results on the larger scale multi-node systems. The performance gains over the older Ivybridge system were very quickly evident and it soon became clear that the teams would all be forced to upgrade their HPC systems in order to remain competitive, which is the nature of Formula 1.

But this upgrade was extremely expensive. This is not what the FIA had been aiming for, but reflects how quickly the HPC industry moves forward with the Formula 1 environment forced to follow suit to remain competitive. Once testing migrated onto the full, multinode systems the full optimisation process could begin. This involved running the same model over a wide range of different set-ups, including options for memory bandwidth per core used and process bindings. The key at this stage was for the group to develop an understanding of the efficiency vs performance of each compute system – ideally each compute chip in each family from Intel and AMD. In reality UniFi and CE were able to use their experience in the industry to limit the testing to the most likely candidates for Formula 1 operations and Boston used its extensive links in the HPC industry to gain access to relevant systems for benchmark testing. Once a small range of AMD and Intel compute chips had been selected, the focus was on understanding how they performed against the Formula 1 regulations. This required repeating the CFD simulation of their Formula 1 car on a range of different HPC sizes and set-ups.

For example, the CFD case will be repeated on the same HPC system but testing the simulation on 48, 96 and 192 cores. It was accepted that the case being run on 96 cores will take slightly longer than half the time of the case on 48 cores and slightly less than half the case being run on 192 cores – so there is an element of inefficiency by running on an increasing number of cores.

"THE NEW METHOD THAT WAS INTRODUCED AT THE START OF THIS YEAR IS A VERY SIMILAR MEASURE TO TFLOPS BUT WITHOUT THE RELIANCE ON FLOPS PER CYCLE"

However, it is in the teams' interest to complete their CFD simulations quickly in order to allow their iterative aerodynamic development programmes to continue as quickly as possible – so it's a trade off and one which was vitally important for the Boston group to understand.

CORE VALUES

The next step was to understand the impact of leaving some of the compute cores dormant, as previously mentioned. This is an approach quite alien to most of the CFD industry (why would you buy compute cores and then not use them?) but something that was already well known to deliver regulatory efficiency in the F1 environment, if you could afford it. Tests were completed leaving a range of the cores dormant in order to give less operational cores per memory channel, and thus increasingly improving the memory bandwidth available to the CFD simulation. The conclusion of this benchmarking study delivered performance gains which would enable the F1 teams to run approximately twice as many CFD simulations per week in 2018 than they had been able to in 2017 (for the same wind tunnel operation). Furthermore, the teams would complete each of these simulations in approximately half the time that was required under the 2017 regulations.

FORMULA 1 SPECIFIC

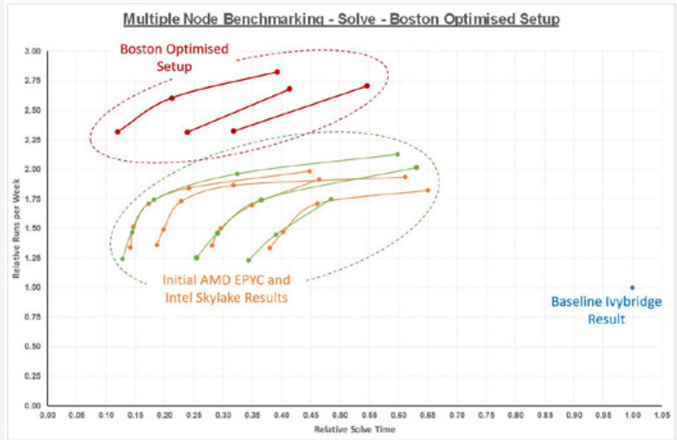
Much of this optimisation is not relevant to the wider CFD industry, but is now considered basic within the Formula 1 teams. The next step was for the Boston group to really exploit the expertise available from the UniFi/CE group. The details of this remain confidential, but it enabled the group to develop solutions which delivered even more performance for the F1 teams, and a further 20 per cent reduction in solve times was extracted from the same CFD set-up, which also increased the CFD throughput by the same 20 per cent.

Finally, as the benchmarking study neared its conclusion Boston worked with AMD to further optimise for the requirements of F1 by increasing the memory bandwidth whilst retaining a relatively low base clock speed. 'AMD EPYC delivers exceptional levels of performance in a number of workloads, including high performance computing CFD applications,' explains Roger Benson, the senior director of the Datacenter Group, EMEA, AMD. 'We are excited to be working with Boston on their automotive engineering focused platforms and improving the efficiency of aerodynamic testing for their customers.'

THE RESULTS

The stated targets of the FIA for this change in regulations was to enable the F1 teams to upgrade from their Fangio and Ivybridge systems to the latest technology available, but without a clear performance pressure to do so, and with the aim of better aligning the F1 industry with the wider CFD industry. Firstly, it is clear that all the F1 teams have upgraded to a new system, with most teams having done so ahead of the regulatory change date of 1 January 2018. So, the first aim has been achieved – the Fangio and Ivybridge systems that the teams were operating are now obsolete.

However, the benchmarking work completed by Boston clearly demonstrates the huge performance advantage available by purchasing a new multi-million pound system, which was not the aim of the new regulations. Furthermore, the impact of the increase in CFD capacity available to the teams under these new regulations only serves to increase the financial pressure on the teams and in particular the pressure to increase headcount within the aerodynamics departments as the CFD capacity



The benchmarking study concluded that teams would gain a huge performance advantage if they purchased a new multimillion pound system because they would have twice the CFD capacity of 2017 – this was not the aim of the regulations.

available increases. Not only have they effectively been required to invest in new HPC architecture in order to remain competitive, but the incentive to adopt future improvements in chip technology has now only increased.

How so? The benchmarking work completed by Boston suggests that teams are now able to complete between 1000 and 1500 CFD simulations per week based on a typical CFD model of around 200 million cells. Teams may elect to 'trade' some of this capacity for larger models (some teams run CFD models approaching one billion cells) or better quality models (transient simulations rather than steady state). But the key point is that the F1 HPC regulations have now given the teams twice as much capacity to play with than in 2017.

STEP CHANGE

Typically each generation of compute chip that is released by AMD/Intel delivers around 10 to 20 per cent improvement in efficiency. Back in 2009 this would give the teams an extra 10 to 20 CFD runs per week, and therefore would not easily justify the large cost in replacing their CFD clusters. In 2018, with the massive increase in capacity, the same 10 to 20 per cent improvement available from each evolution of compute chip technology is 100 to 200 runs – that is the same as the total capacity of the systems in 2009.

Is this a bad thing? Arguably not. HPC systems are much cheaper now than they were back in 2009. The FIA focus remains on reducing wind tunnel reliance and delivering greater CFD capacity in exchange, and the current regulations deliver that. However, does it help to level the playing field between the high budget teams and the low budget teams? Does it help to encourage new teams into the sport? And does it make the working practices within the Formula 1 aero departments more aligned to the wider CFD industry?

With AMD releasing its second generation of EPYC chip in 2019, the reaction of the teams will be interesting. Will they all upgrade immediately? Or will the well-funded teams take the opportunity to get a performance advantage from the new technology that the smaller teams cannot afford?

Boston, UniFi and CE continue to develop their partnership with a focus on the F1, motorsport and automotive industries across all CFD codes and working practices. For more information visit the websites at: www.boston.co.uk; www.unifimotorsport.com; www.computationalengineering.co.uk

SINCE THIS ARTICLE WAS PUBLISHED AMD HAS RELEASED IT'S SECOND GENERATION OF EYPC™ CHIP WHICH IS NOW AVAILABLE FOR TESTING IN BOSTONLABS