WHITE PAPER

SCALABLE PERFORMANCE WITH THE HARLEQUIN RIP®

GLOBAL GRAPHICS® software

Introduction

The Harlequin RIP® from Global Graphics Software is a world-leading RIP for processing Page Description Languages (PDLs) including PostScript®, PDF and XPS. It was first released as a shipping product in 1988 and has been proven in production print ever since, including for digital print since 2002. Over its lifetime it has evolved in many ways; in fact it's been virtually re-written on several occasions, each time emerging even faster and more efficient.

The impetus for each new step in its development has come from many sources. Three of the most important, at least for performance, have been, and continue to be:

- The increasing speed of output devices, continually raising the bar for raster delivery. To take an example, the HP T1190 inkjet press that is driven by the Harlequin RIP today requires approximately 45GB of raster data every second to keep it running at engine speed, 305m per minute.
- The increasing complexity of documents to be printed, especially the inclusion of live PDF transparency for high volume digital press production. Correct rendering of live transparency greatly increases the amount of processing required per page.
- The migration from single-core, single-CPU computers, through machines with multiple single-core CPUs to today's model where high end servers contain multiple CPUs, each comprising many cores, and a digital front end for the fastest presses contains many servers.

Any software RIP today is probably running on a computer with at least four cores, even if it's being used to drive a device needing a relatively low data rate, such as a proofing printer, CTP platesetter or a laser printer used for light production. If the computer has been specifically selected to deliver high speed RIPing then it will have many more.

This White Paper looks at how the scalability of the Harlequin RIP® enables rapid implementation of cost-effective solutions and maximum throughput on digital print devices, from light production to ultra-high volume.

What does a RIP have to do?

RIPing a job comprises several steps (see 'RIP phases' below), some of which can be processed by multiple threads in parallel. Conceptually each major thread, performing a task that requires significant CPU time, can be thought of as running on a single core, although thread management schemes may mean that the reality is more complicated. Thus a RIP on an eight core computer might be configured to spawn seven major threads for heavyweight processing, making use of seven cores (leaving one for the operating system)*. Lightweight management threads that don't use much CPU time typically share a core with major processing threads.

RIP phases

The process of RIPing a page requires several steps to be performed in order, regardless of whether that page is submitted as PostScript, PDF or any other page description language

Interpreting: the page description language to be RIPed is read and decoded into an internal database of graphical elements that must be placed on the page. Each may be an image, a character of text (including font, size, color etc), a fill or stroke etc. This database is referred to as a display list.

Compositing: The display list is pre-processed to apply any live transparency that may be in the job. This phase is only required for any pages in PDF and XPS jobs that use live transparency; it's not required for PostScript language pages because those cannot include live transparency.

Rendering: The display list is processed to convert every graphical element into the appropriate pattern of pixels to form the output raster. The term 'rendering' is sometimes used specifically for this part of the overall processing, and sometimes to describe the whole of the RIPing process. It's only used it in the first sense in this document.

Output: the raster produced by the rendering process is sent to the marking engine in the output device, whether it's exposing a plate, a drum for marking with toner, an inkjet head or any other technology.

Sometimes this step is completely decoupled from the RIP, perhaps because plate images are stored as TIFF™ files and then sent to a CTP platesetter later, or because a near-line or off-line RIP is used for a digital press. In other environments the output stage is tightly coupled with rendering.

RIPing often includes a number of additional processes; in Harlequin RIPs:

- In-RIP imposition is performed during interpretation
- Color management (Harlequin ColorPro®) and calibration are applied during interpretation or compositing, depending on configuration and job content
- Screening is applied during rendering or after the Harlequin RIP has delivered unscreened raster data if screening is being applied post-RIP, when Global Graphics' ScreenPro and PrintFlat technologies are being used, for example.

These are all important processes in many print workflows, but they don't materially affect the primary flow of data through the main four work steps, so they are omitted from the descriptions in this white paper to avoid making diagrams too complex.

*Hyper-threading has been ignored here in the interests of simplicity.

Interpreting is necessarily a serial process, because a single stream of page description must be read in order so that the objects can be positioned correctly relative to each other.

Transparency compositing and rendering can be processed by several threads in parallel, by dividing the whole page into bands that can be processed independently. The rich use of multi-threaded techniques in Harlequin Host Renderer 2 is illustrated in FIG 1. In this diagram each row of colored blocks represents a series of software threads.

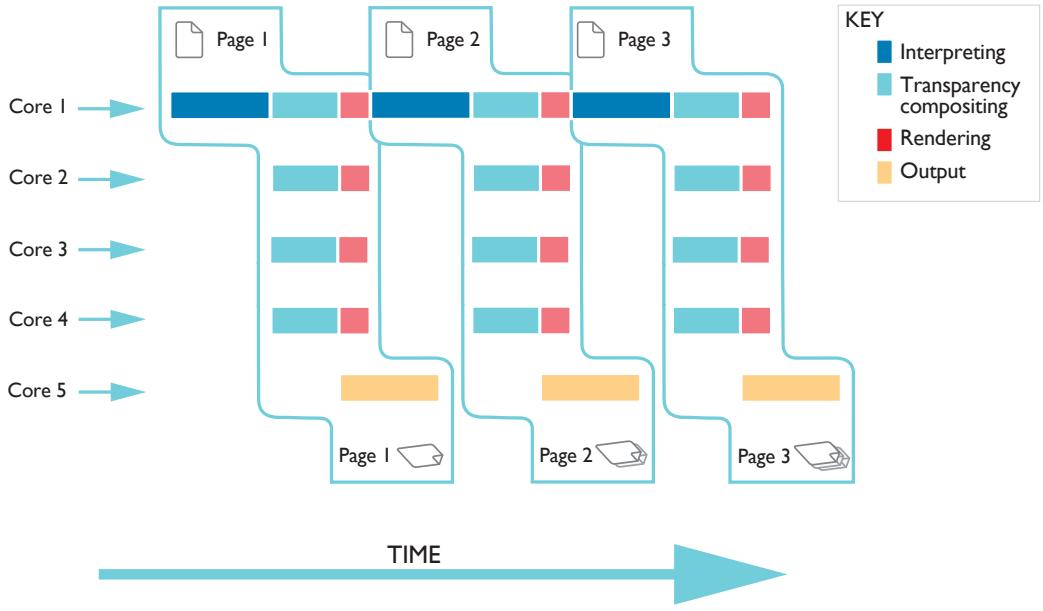

FIG 1 - Very simple Multi-threading in Harlequin Host Renderer. Multiple threads are used for transparency compositing, rendering and outputting the pages of a job at once, but interpreting introduces periods when only a single-thread is active at once.

The number of threads that can be allocated to the RIP can be configured; this diagram shows 5 concurrent threads in action, which would be suitable for a six core computer, leaving one more core for the operating system and other ancillary tasks.

A typical performance monitor trace for the Harlequin RIP processing a job using multi-threaded compositing and rendering as shown in FIG 1 looks like FIG 2, using all of the cores for about half of the time, while each new page is composited and rendered, but interpretation still only uses a single thread.

This trace shows that several cores are idle while pages are being interpreted, which means that there is still more performance to be squeezed out of the computer.

FIG 2 – processor activity trace using multi-threaded compositing and rendering, but not Harlequin Parallel Pages™.

FIG 3 – processor activity trace using multi-threaded compositing and rendering, in combination with Harlequin Parallel Pages is 1.3 times faster.

Advanced use of multiple cores

This inefficiency led to the development of Harlequin Parallel Pages™. It decouples interpretation from compositing and rendering, allowing one thread to interpret page two of a job while the other threads are still compositing, rendering and outputting page one. There's still a period while the RIP is interpreting the first page of a job where only one core is fully utilized, but for later pages all cores are used virtually all of the time to maximize throughput.

The resulting performance monitor trace therefore looks like FIG 3, which shows that all of the cores in the processor are working hard for a much larger proportion of the time. As a result, the total time for the job is significantly reduced. FIG 2 and FIG 3 are at the same scale and for the same job, using configurations that only differ in the use of Harlequin Parallel Pages. The length of the trace represents the processing time for the whole job. As these traces show, Harlequin Parallel Pages allows the Harlequin RIP to process this job over 1.3 times faster.

The same increased utilization of the processor cores and compression of the processing time can be seen in FIG 4, which extends the model shown in FIG 1 to also include Harlequin Parallel Pages.

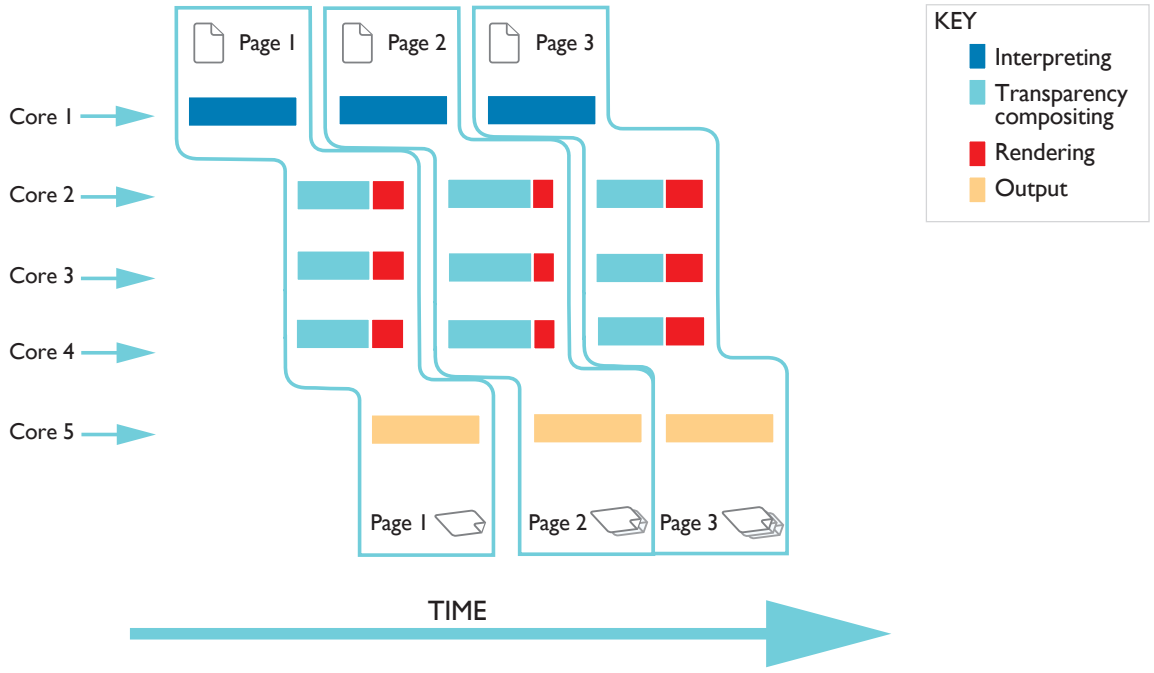

FIG 4 - Multi-threading using Harlequin Parallel Pages™. Multiple threads are used to perform the various steps of RIPing and outputting the pages of a job at once; there are no longer any periods when only a single thread is active.

RIP farms

However efficient a single RIP can be made, there comes a point at which it can't deliver a data rate high enough for today's fastest digital presses in real time. Global Graphics has therefore provided a framework for configuring, controlling and tracking multiple instances of the Harlequin RIP. It's named for what it does: Harlequin Scalable RIP.

The Harlequin Scalable RIP can be used in two different modes: Single-server Scalable RIP, for running multiple RIPs on a single computer, and Multi-server Scalable RIP for running multiple RIPs across multiple computers.

In both cases the Scalable RIP handles replication of configurations between RIPs, automatically splits PDF files across multiple RIPs and manages all metadata about rasters delivered. The job is simply submitted to the Scalable RIP with a selected configuration in a single transaction.

This means that there are three significant steps in Harlequin scalability:

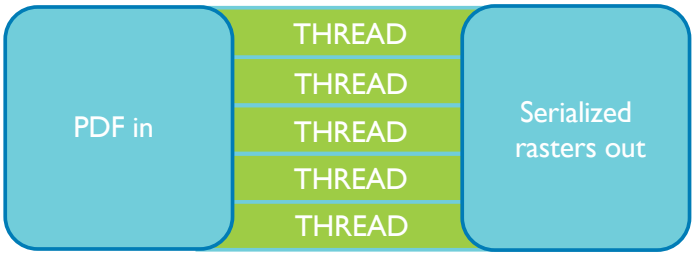

FIG 5 - A single instance of Harlequin, using multiple cores

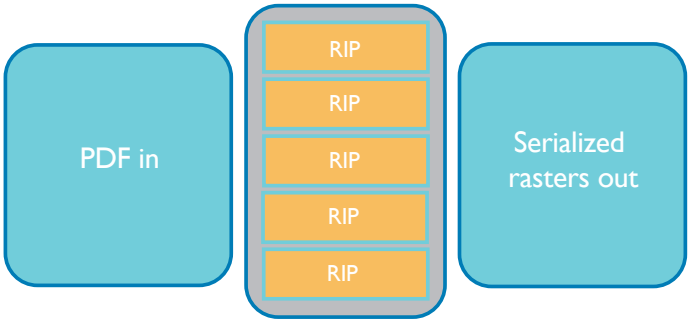

FIG 6 - Single-server Harlequin Scalable RIP

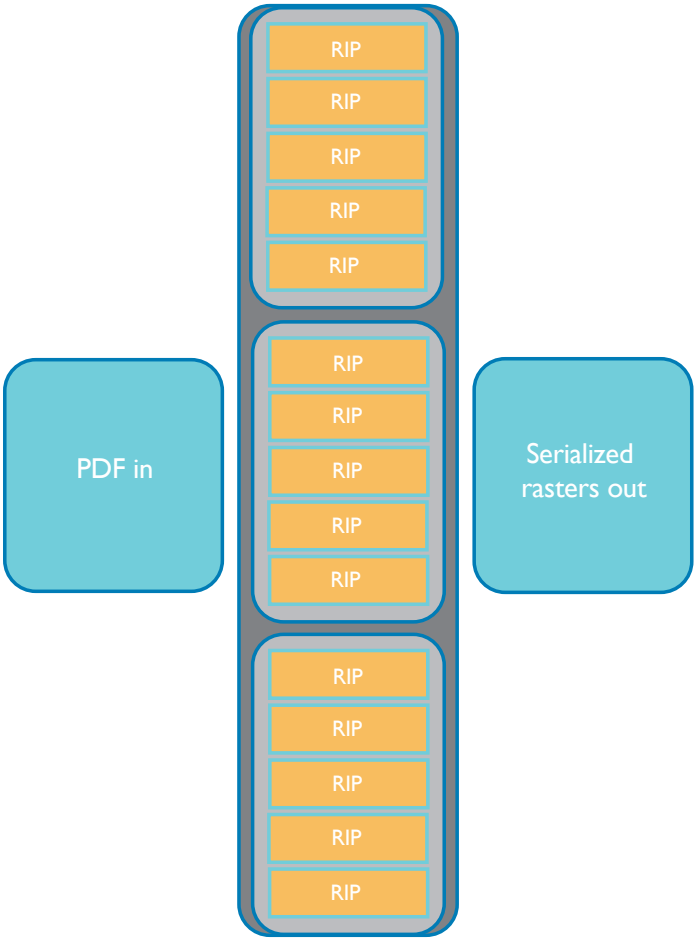

FIG 7 - Multi-server Harlequin Scalable RIP

Handling the raster delivery from a farm of RIPs will always be a little more complex than receiving the rasters from a single RIP. When only one RIP is rendering, the rasters will always be finished in the order that the DFE code requested them.

But when you have multiple RIPs working on 'chunks' of pages in parallel it's highly likely that pages will sometimes be completed out of order, if only because some pages will be more complex than others, and therefore take longer to process. And the first pages of one job may also be completed before the last few of the previous one, especially if you're processing a large number of small jobs across a lot of RIPs.

The Harlequin Scalable RIP therefore includes raster management functionality. This does not constrain how and where you store the rasters within your DFE (whether that's on disk, in shared memory, etc), but it assists with managing metadata about those rasters, and in serializing them so that downstream processes can be provided with all the information about those rasters in the order that they need them.

Classes of output device

Most of us wouldn't try to move house in a racing car, or take the kids to school in a 40 ton lorry; we use a vehicle that was designed to deliver the range of motoring requirements that we expect to encounter in our daily lives. In the same way, a significant range of variations on Harlequin configurations can be applied to different classes of device.

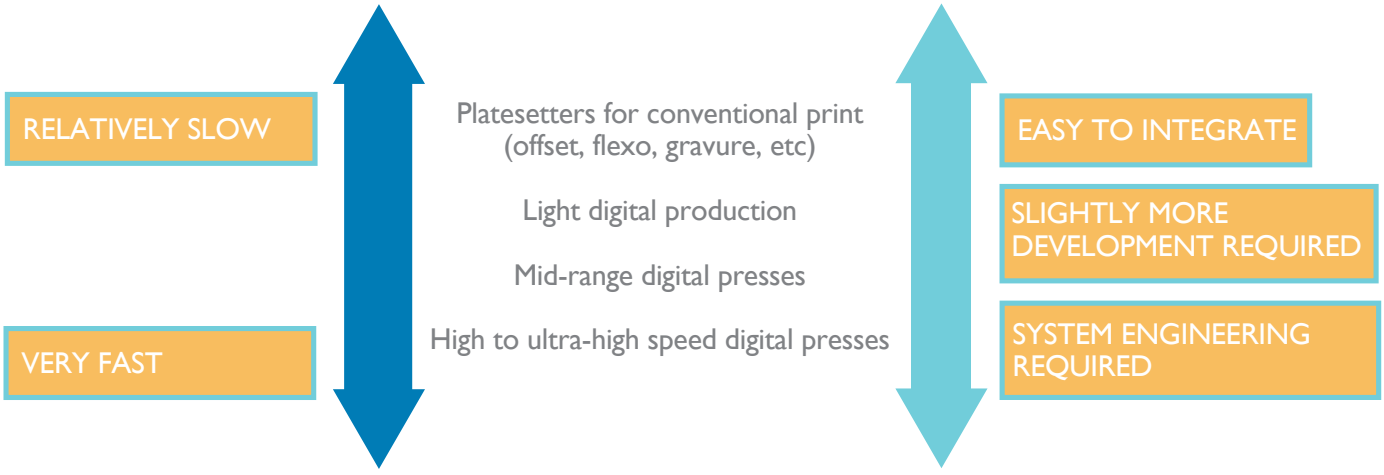

FIG 8

This diagram shows clearly that the integration development necessary increases with the speed of the device, especially when you're trying to get the maximum performance out of your hardware. This is exactly why Global Graphics have developed the Harlequin Scalable RIP: to reduce the integration effort required to build a DFE to drive the high-speed digital presses that are relatively common these days.

At the simplest level, it takes minutes to start using a Single-server Harlequin Scalable RIP to make rasters on disk. And if you already have the code to deliver rasters to the driver for your electronics, it won't take long for your development team to achieve their first print. A tighter integration is often advantageous for better error handling and throughput, and that will take a little longer, of course.

But rather than characterizing presses simply by categories like these, it's better to calculate the data rate that any particular device requires, using the details shown in the box on the next page.

Calculating data rates

Over the years at Global Graphics Software, we've found that the best guidance in sizing a Harlequin solution for a digital press comes from a relatively simple calculation of the data rate required on the output side.

For each class of press, multiply the items listed below together, and then divide by 60 million to give the data rate in megabytes (MB) per second.

You may notice that these calculations are for a data rate to deliver un-screened 8 bits per pixel (contone) rasters. This has proven to be a better metric for estimating RIP requirements than making adjustments based on whether the RIP is applying halftones. In practice Harlequin will run at about the same speed for 8-bit contone and for 1-bit halftone output because the extra work of halftoning is offset by the reduced volume of raster data to move around. Multi-level halftones delivered in 2-bit or 4-bit rasters take a little bit longer, but not enough to need to be considered here.

You can also use the sheet-fed calculation below for conventional print plate setters.

For a single-pass, web-fed inkjet press, multiply these items together:

- The resolution across the press, in dots per inch.
- The resolution along the press, in dots per inch.
- The maximum printable width, in inches.
- The speed of the substrate through the press, in inches per minute.

Alternatively, use meters for the width and meters per minute for the speed, then multiply your final value by 1550 (the number of square inches in a square meter).

- The number of independent colorants. (Multiple print bars used for higher speed output should not be counted here, as they're accounted for under resolution and speed above. Multiple bars for redundancy should also not be included because they don't affect the rendering speed required).
- If it's printing on both sides of the substrate at the same time, multiply by 2

For a sheet-fed press, multiply these items:

- The resolution across the press, in dots per inch.
- The resolution along the press, in dots per inch.
- The number of Letter or A4 pages per minute. (This will include any effect from printing on both sides of the substrate at once).
- The number of independent colorants
- 95, the average number of square inches for Letter and A4 pages

For presses quoted in area per minute, multiply these items:

- The resolution across the press, in dots per inch.
- The resolution along the press, in dots per inch.
- The area printed per minute, in square inches (or square feet * 144, or square meters * 1550)
- The number of independent colorants (light and light-light inks count as separate colorants from the full density inks, so a press with CMYK and light CMK would count as having seven colorants)

And for engineers focused on inkjet heads, multiply these items:

- The number of heads (totaled across all colorants)
- The number of active nozzles per head (don't include nozzles that are not being used, e.g. in stitching zones)
- The head frequency (in kHz)
- 0.06, to yield bytes per minute

The data rate number can be used to make a first approximation of which class of Harlequin integration you should be considering.

- Up to 250MB/s: can be done with a single Harlequin RIP using multi-threading in that RIP
- Up to 1GB/s: use multiple RIPs in Single-server Harlequin Scalable RIP
- Over 1GB/s: use multiple RIPs in Multi-server Harlequin Scalable RIP

These numbers indicate the data rate that the RIP needs to provide when every copy of the output is different. The value may need to be adjusted for other scenarios:

- If you're printing the same raster many times, the RIP data rate should be reduced in proportion
- If you're printing variable data print jobs with significant re-use of graphical elements between copies, then Harlequin VariData can be used to accelerate processing. This effect is already factored into the recommendations above.

The complexity of the jobs you're rendering will also have an impact.

Transactional or industrial labelling jobs, for example, tend to be very simple, with virtually no live PDF transparency and relatively low image coverage. They are therefore typically fast to render. If your data rate calculation puts you just above a threshold in the list above, you may be able to take one step down to a simpler system.

On the other hand, jobs such as complex marketing designs or photobooks are very image-heavy and tend to use a lot of live transparency. If your data rate is just below a threshold on the list above, you may need to step up to a higher level of system.

Be careful when making adjustments. If you do so you may choose either to build and support multiple variations of your DFE, to support different classes of print site, or to design a single model of DFE that can cope with the great majority of your customers. Building a single model certainly reduces development, test and support costs, and may reduce your average bill of materials. But doing that also tends to mean that you will need to base your design on the raw, "every copy different", data rate requirements.

Our experience has also been that the complexity of jobs in any particular sector is increasing over time, and the run lengths that people will want to print are shortening. Designing for current expectations may give you an under-powered solution in a few years' time, maybe even by the time you ship your first digital press. Moore's law, that computers will continue to deliver higher and higher performance at about the same price point, will cancel out some of that effect, but probably not all of it.

And finally, the recommendations above implicitly assume that a suitable computer configuration is used. You won't achieve 1GB/s output from multiple RIPs on a computer with one, four-core CPU, for example. Key aspects of hardware affecting speed are: number of cores, CPU clock speed, disk space available, RAM available, disk read and write speed, band-width to memory, L2 and L3 cache sizes on the CPU and (especially for multi-server configurations) network speed and bandwidth.

Interaction with Harlequin VariData™

Harlequin RIPs use a feature call Harlequin VariData to greatly accelerate the processing of variable data jobs in PDF. It works by identifying groups of graphical elements that are used multiple times and by only rendering each group once. As the pages are processed the shared elements are not rendered again. Harlequin VariData can be configured in two ways.

- In internal mode the pre-rendered rasters for shared elements are merged into the page as it's processed. The RIP delivers complete page rasters, making it an easy way for new digital press vendors to accelerate variable data print.
- In external mode the shared rasters are delivered to an external cache and are merged with single-use rasters post-RIP, allowing press vendors with their own composition technology to maximize RIP throughput.

In both situations the RIP performs more separate compositing and rendering operations when using Harlequin VariData, but each individual operation is smaller. The same multi-threading capability is applied to each of these operations as it would be without Harlequin VariData

Why is scalability important?

A lot of time and effort has been invested in adding, updating and extending multi-threaded processing and cooperative processing between multiple RIPs in the Harlequin RIP, but raw speed is obviously not the real goal of the exercise. What really matters is providing a RIP that can deliver maximum throughput on your output device.

In parallel with the multi-threading development, therefore, a huge amount of attention has been paid to avoiding unnecessary processing and to optimizing code. As a result the RIP requires fewer CPU cycles to process each page, and can spread those cycles over more threads, for a win-win result.

With the right architecture and a large enough number of RIPs virtually any conceivable digital press can be driven at engine speed with typical print jobs, but the resulting DFE could be uneconomical in comparison to the cost of the press. By making every RIP faster and by maximizing the utilization of the hardware Harlequin RIPs can reduce the total bill of materials for a DFE by a significant amount. If you don't need as many RIPs, a DFE may be built using fewer computers (or a lower cost computer if it only uses one). Reducing the number of computers may also reduce costs for operating systems, anti-virus software etc, as well as allowing a smaller footprint for the DFE. And finally print sites can also save on power and cooling as a result; a solution that's cost effective for the press vendor is also greener for the printing company!

About the author

Martin Bailey, CTO, Global Graphics Software

Martin Bailey works to analyze and understand current and future needs for workflows across many sectors of print. This enables him to guide Global Graphics' industry-leading printing technology. He represents Global Graphics on a number of industry bodies and standards committees including acting as the primary UK expert on the committees working on PDF, PDF/X and PDF/VT.

Martin has over 30 years of experience building, using, supporting and improving products for processing digital documents and the print industry in technical support, product management and programming as well as in consulting, and production environments.

About Global Graphics Software

Global Graphics Software http://www.globalgraphics.com is a leading developer of platforms for digital printing, including the Harlequin RIP®, ScreenPro, Fundamentals and Mako. Customers include HP, Canon, Durst, Roland, Kodak and Agfa. The roots of the company go back to 1986 and to the iconic university town of Cambridge, and, today the majority of the R&D team is still based near here. Global Graphics Software is a subsidiary of Global Graphics PLC (Euronext: GLOG).

GLOBAL GRAPHICS® software

April 2019 v2

Contact us.
info@globalgraphics.com

www.globalgraphics.com

Global Graphics Software Inc.
5996 Clark Center Avenue
Sarasota, FL 34238
United States of America
Tel: +1(941)925-1303

Global Graphics Software Ltd
Building 2030
Cambourne Business Park
Cambourne, Cambridge
CB23 6DW UK
Tel: +44 (0)1954 283100

Global Graphics KK
610 AIOS Nagatacho Bldg.
2-17-17 Nagatacho, Chiyoda-ku,
Tokyo 100-0014
Japan
Tel: +81-3-6273-3198