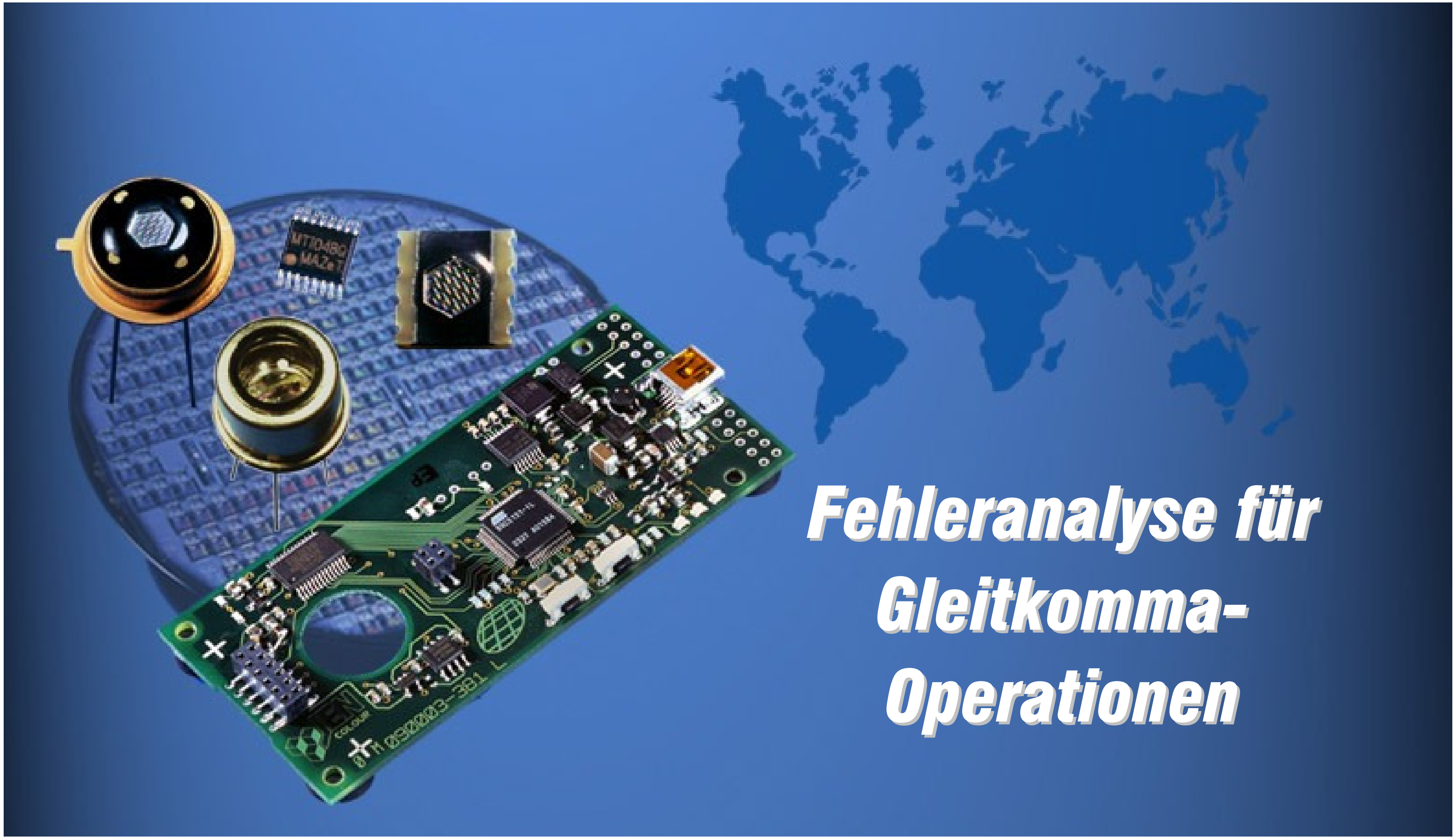
Farbworkshop ‘07

JENCOLOR

Sensor Design
and Manufacturing Services



Fehleranalyse für Gleitkomma-Operationen

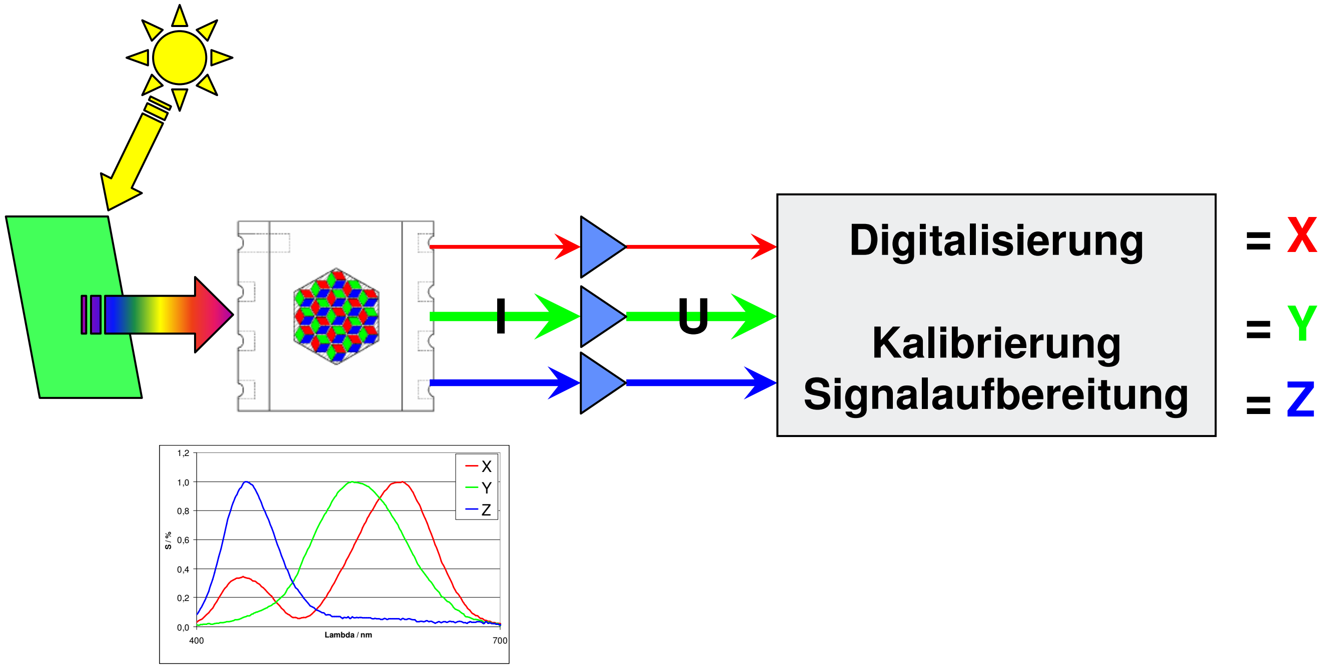
Einleitung

Sensor Design
and Manufacturing Services

- Kurzeinführung ColorLib
- Fließkommanotation reeller Zahlen auf Computer
- Grundrechenarten
- Fehlerfortpflanzung
- Vergleich der Darstellungsformen
- Zusammenfassung

Motivation

- **Hoher Aufwand der Programmierung farbmetrischer Operationen**
- **Keine einfache „Plug and Play“ Lösung**



Aufgaben der ColorLib

- **Funktionsbibliotheken für:**
  - Digitalisierung
  - Kalibrierung
  - Messwertkorrektur
  - Farbmetrische Berechnungen
  - Zuordnungen der Daten

Umsetzung

JENCOLOR

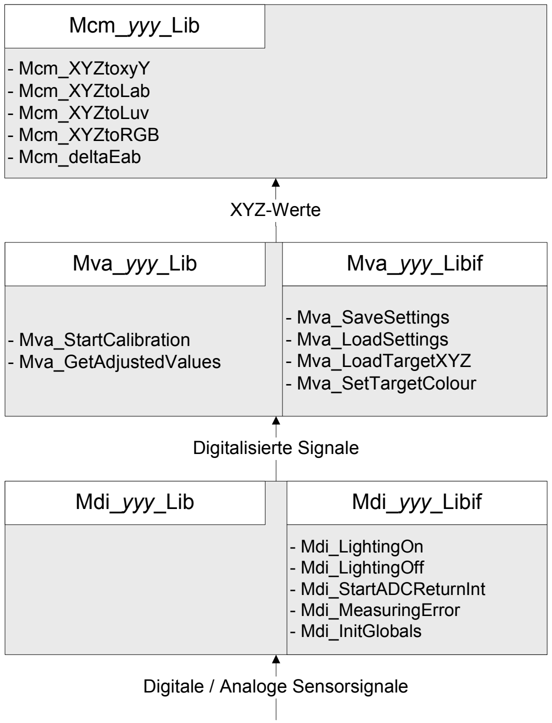
Sensor Design
and Manufacturing Services

- Zielsysteme: PC, 32-, 16-bit Mikrocontroller
  (8-bit Mikrocontroller)
- Programmierung in ANSI C
- Kalibrierung und Auslesen der Sensoren jeweils mit einem Funktionsaufruf
- Einbindung zielsystemabhängiger Funktionen über Callback-Funktionen
- Keine Einbindung von Standardbibliotheken

Lösung

Sensor Design
and Manufacturing Services

- Nahezu simultane Auswertung mehrerer True Color Sensoren
- Ausgabe korrigierter Messwerte in einem Array
- Einbindung einer Software Gleitkomma-Bibliothek



Fließkommadarstellung

Sensor Design
and Manufacturing Services

- **Wissenschaftliche Notation**
- **Mathematische Beschreibung einer Zahl z:**
  - Vorzeichen vz
  - Mantisse m
  - Basis b
  - Exponent x

$$z = (-1)^{vz} \cdot m \cdot b^{x}$$

Fließkommadarstellung

Sensor Design
and Manufacturing Services

- **Wahl des Exponenten beschreibt den Wertebereich**
- **Basis kann frei gewählt werden**
  - durch IEEE 754 hat sich die Basis 2 auf Computern durchgesetzt
- **Mantisse enthält die grundlegenden Ziffern einer Fließkommazahl**
  - Je mehr Ziffern, desto höhere Genauigkeit

Standard - IEEE 754

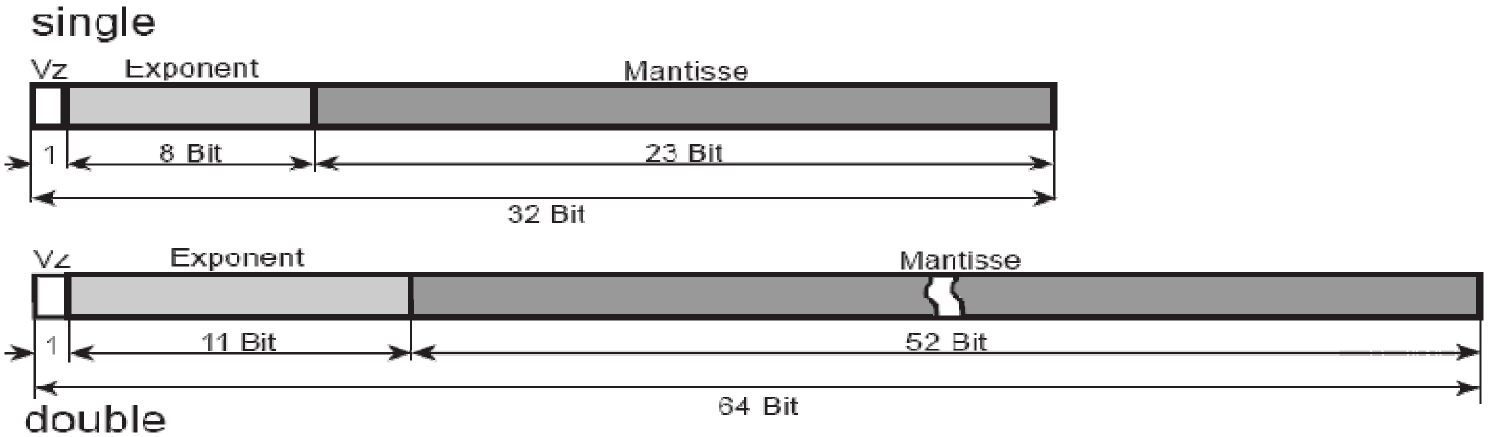
JENCOLOR
Sensor Design and Manufacturing Services

- **Es existiert eine Vielzahl an unterschiedlichen Formaten**
  - Anordnung der Daten
  - Wahl der Parameter zur Darstellung (Mantisse, Exponent)
- **IEEE 754 definiert die Eigenschaften für binäre Fließkommazahlen auf dem Computer**
  - Rundungsmodi, Sonderfälle, Datenformat

IEEE 754

- **Die Kommission definiert 4 verschiedene Grundformate:**
  - single
  - double
  - single extended
  - double extended
- **Abspeicherung in der Form**
  - 1 Bit Vorzeichen
  - m Bit Exponent
  - n Bit Mantisse

	Min	**Max**
single	$\approx 1{,}2 \cdot 10^{-38}$	$\approx 3{,}4 \cdot 10^{38}$
double	$\approx 2{,}2 \cdot 10^{-308}$	$\approx 1{,}8 \cdot 10^{308}$
ColorLib	$\approx 2{,}9 \cdot 10^{-39}$	$\approx 1{,}7 \cdot 10^{38}$

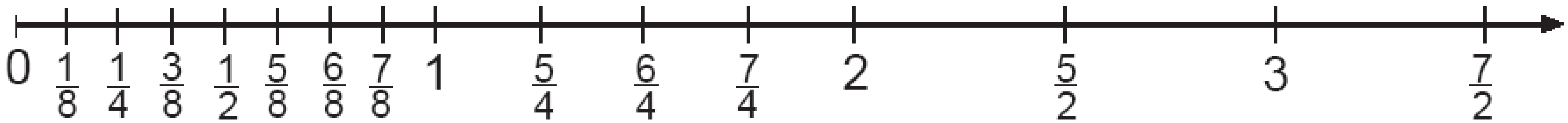


Darstellbare Zahlen

- **Durch das separate Vorzeichen sind die Zahlen symmetrisch um die 0 angeordnet**
- **Beispiel:**

$$1{,}0_{10} = 1.00_2 \cdot 2^0$$

$$0{,}125_{10} = 0.01_2 \cdot 2^{-1}$$



Arithmetische Operationen von Fließkommazahlen - Addition

JENCOLOR
Sensor Design
and Manufacturing Services

$$\left.\begin{aligned} X &= S_x \cdot M_x \cdot 2^{E_x} \\ Y &= S_y \cdot M_y \cdot 2^{E_y} \end{aligned}\right\} \Rightarrow Z = S_x \cdot M_x \cdot 2^{E_x} + S_y \cdot M_y \cdot 2^{E_y}$$

$$\Rightarrow \left(S_x \cdot M_x + S_y \cdot M_y \cdot 2^{E_y - E_x}\right) \cdot 2^{E_x}$$

$$\Rightarrow \left(S_x \cdot M_x + S_y \cdot m_y\right) \cdot 2^{E_x}$$

$$\Rightarrow S_z m_z \cdot 2^{E_x} \Rightarrow S_z M_z \cdot 2^{E_z}$$

- **Anpassen des Exponenten E_y an E_x**
  - Verschiebung der Mantissen
- **Festkommaaddition / -subtraktion der Mantissen**
- **Normalisierung**

Problem der Auslöschung während der Subtraktion

JENCOLOR

Sensor Design
and Manufacturing Services

- **Aufgrund des begrenzten Speicherplatzes werden signifikante Stellen ausgelöscht**
- **Nur umgehbar durch komplexe Umformungen von Termen (Subtraktionen eliminieren)**

$$
\begin{aligned}
&1{,}1111 \mid xxxx \cdot 2^{0} \\
-&1{,}1110 \mid xxxx \cdot 2^{0} \\
=&00001 \mid xxxx \cdot 2^{0} = 1,\underbrace{0000}_{Inkorrekt} \cdot 2^{-4}
\end{aligned}
$$

Arithmetische Operationen von Fließkommazahlen - Multiplikation

JENCOLOR

Sensor Design
and Manufacturing Services

$$\left.\begin{array}{l} X = S_x \cdot M_x \cdot 2^{E_x} \\ Y = S_y \cdot M_y \cdot 2^{E_y} \end{array}\right\} \Rightarrow Z = S_x \cdot M_x \cdot 2^{E_x} \cdot S_y \cdot M_y \cdot 2^{E_y}$$

$$\Rightarrow \left(S_x \cdot M_x \cdot S_y \cdot M_y\right) \cdot 2^{E_x + E_y}$$

$$\Rightarrow S_z m_z \cdot 2^{E_z} \Rightarrow S_z M_z \cdot 2^{E_z}$$

- **Exponenten addieren**
- **Festkommamultiplikation der Mantissen**
- **Normalisierung**

Multiplikation der Mantissen

Sensor Design
and Manufacturing Services

- **Problem Überlauf: Multiplikation größerer Zahlen**
- **Lösung 1: Auf Binärebene die Multiplikation durchführen**
  - Nachteil: relativ langsam, da die Ermittlung Bit für Bit durchgeführt werden muss
- **Lösung 2: Aufteilen der Werte (Bsp.: 2x32 Bit)**

Arithmetische Operationen von Fließkommazahlen - Division

JENCOLOR

Sensor Design
and Manufacturing Services

$$\left.\begin{aligned} X &= S_x \cdot M_x \cdot 2^{E_x} \\ Y &= S_y \cdot M_y \cdot 2^{E_y} \end{aligned}\right\} \Rightarrow Z = \frac{S_x \cdot M_x \cdot 2^{E_x}}{S_y \cdot M_y \cdot 2^{E_y}} \Rightarrow \frac{S_x \cdot M_x}{S_y \cdot M_y} \cdot 2^{E_x - E_y}$$

$$\Rightarrow S_z m_z \cdot 2^{E_z} \Rightarrow S_z M_z \cdot 2^{E_z}$$

- **Exponenten subtrahieren**
- **Festkommadivision der Mantissen**
- **Normalisierung**

Division der Mantissen

Sensor Design
and Manufacturing Services

- Problem ganzzahlige Division von Zahlen mit gleicher Wortbreite
- 32 Bit / 32 Bit = 1 Bit

Arithmetische Operationen von Fließkommazahlen – n-te Wurzel

Sensor Design
and Manufacturing Services

$$Y = S_y \cdot M_y \cdot 2^{E_y}$$

$$\Rightarrow Z = \sqrt[n]{Y} = Y^{\frac{1}{n}}$$

$$\Rightarrow Z = \sqrt[n]{S_y \cdot M_y}\, 2^{\frac{E_y}{n}}$$

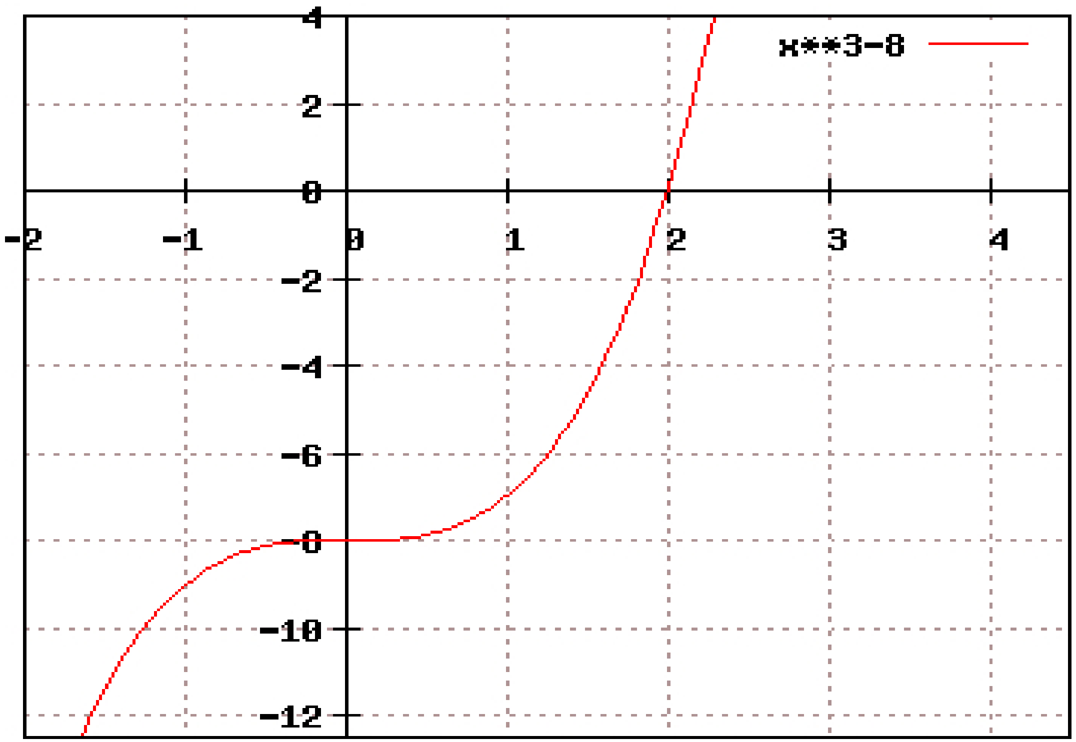
$$\Rightarrow Z = S_z m_z \cdot 2^{E_z} \Rightarrow S_z M_z \cdot 2^{E_z}$$

- **Exponent durch n teilen**
  - Wenn Exponent nicht ganzzahlig teilbar, dann Mantisse verschieben um maximal n-1 Stellen
- **n-te Wurzel aus der Mantisse ziehen**
- **Normalisierung**

Darstellung der n-ten Wurzel

$$f(x) = x^n - a$$

$$\sqrt[n]{a} = x$$



Wie kann die n-te Wurzel bestimmt werden?

- **Lösung über iterative Verfahren, die die Nullstelle der Funktion f(x) approximieren.**
- **Nach Newton:**

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) \qquad x_{n+1} = \frac{1}{3}\left(2x_n + \frac{a}{x_n}\right)$$

Größt-Fehlerfortpflanzung in den arithmetischen Operationen

JENCOLOR
Sensor Design
and Manufacturing Services

- **Addition**

$$\varepsilon_{x\pm y} = \frac{x}{x\pm y}\varepsilon_x \pm \frac{x}{x\pm y}\varepsilon_y$$

- **Multiplikation**

$$\varepsilon_{x\cdot y} = \varepsilon_x + \varepsilon_y + \varepsilon_x\varepsilon_y$$

- **Division**

$$\varepsilon_{\frac{x}{y}} = \frac{\varepsilon_x + \varepsilon_y}{1+\varepsilon_y}$$

Ergebnis:
Vergleich bezogen auf Farbmetrik

JENCOLOR
Sensor Design
and Manufacturing Services

▶ **Berechnung in einzelne Farbräume mit vorheriger Kalibrierung:**

	ColorLib	single	double
Speicher	48bit	32bit	64bit
XYZ	3,43E-07	8,78E-05	1,64E-13
xy	6,86E-07	1,76E-04	3,27E-13
L*	9,44E-07	1,12E-04	3,68E-11
a*	1,88E-06	2,24E-04	7,36E-11
b*	1,88E-06	2,24E-04	7,36E-11
u*	1,63E-06	2,88E-04	3,72E-11
v*	1,63E-06	2,88E-04	3,72E-11
ΔEab	3,77E-06	4,49E-04	1,47E-10

→ **ColorLib liegt in Bezug auf Speicherbedarf und Genauigkeit zwischen single und double.**

Zusammenfassung

JENCOLOR

Sensor Design
and Manufacturing Services

- **ColorLib und Gleitkommabibliothek aktuell als beta-Version für PC vorhanden**
- **Portierbarkeit von PC auf Mikrocontroller gegeben**
- **Optimaler Kompromiss aus Genauigkeit, Rechengeschwindigkeit und Speicherbedarf angepasst für farbmetrische Berechnungen auf Mikrocontroller**
- **Kostengünstiger als zusätzliche Floating Point Unit (FPU)**

Vielen Dank...

Sensor Design
and Manufacturing Services